# Expertise in Applied Face Matching: Training, Forensic Examiners, Super Matchers and Algorithms

## Thesis

# *Expertise in applied face matching: training, forensic examiners, super matchers and algorithms*

Reuben Edwin Leigh Moreton

Submitted to the Open University, Faculty of Arts and Social Sciences

In partial fulfilment of the requirements for the degree of

Doctor of Philosophy

March 2021

## *Abstract*

Face matching is widely used in applied settings, including policing and border control, to identify persons of interest, where the consequences of an incorrect decision can have profound consequences. It is, therefore, of paramount importance that applied face-matching systems are accurate and reliable. However, humans are generally poor at matching face of people they don't know, with large individual differences in accuracy. The aim of this thesis was to evaluate different sources of face-matching expertise (training, forensic face examination, superior face matchers and algorithms) and provide recommendations for how to improve face-matching performance in applied settings.

Study one presents a survey of face-matching training, providing insights into the diverse and inconsistent approaches organisations use to train face-matching operators. The second study evaluates a two-day professional face-matching training course, demonstrating the limitations of short courses and the risk of introducing a match bias in low performers. In study three the perceptual skill of superior face matchers and forensic face examiners were compared, showing that by combining the selection of high performers with a wisdom of crowds approach, comparable levels of performance to trained examiners can be achieved in quick-decision face matching. Study four investigated the fusion of human face-matching decisions and algorithm similarity scores for faces that were challenging to humans and to the algorithm, highlighting the effectiveness of fusion in improving face-matching performance. Study five compared the operational accuracy of individual examiners and examiner teams on a face-matching task. Teams achieved higher levels of performance than individuals, with performance improving for both groups after fusion with a facial recognition algorithm.

The thesis concludes with a discussion of how different sources of face-matching expertise can be used and combined in applied face-matching systems, and highlights areas for further research that would benefit the applied face-matching community.

## *Acknowledgements*

Firstly, I would like to sincerely thank my supervisors, Dr Catriona Havard, Dr Ailsa Strathie and Prof Graham Pike, for their continued support, guidance, expertise and enthusiasm throughout my PhD.

This PhD would not have been possible without the help of the police personnel who contributed as participants, particularly my friends and colleagues at the Metropolitan Police Service, thank you all. My sincere thanks also to the forensic face examiner community, in particular the European Network of Forensic Science Institutes Digital Imaging Working Group for your contributions to this thesis.

A special thanks to Cosima Calder from Qumodo for your help with all things relating to facial recognition algorithms, Dr Jonathon Phillips at the National Institute for Standards and Technology for advice on the fusion analysis and Dr Alice Towler, Dr David White and Prof Richard Kemp at the University of New South Wales for your continued support and helpful discussions over the years.

I would also like to acknowledge the support of the Open University's Faculty of Arts and Social Science and the Centre of Policing Research and Learning for resources and support provided throughout my PhD, including funding to attend conferences here in the UK and abroad.

Thank you Louisa, always, and to my son Eli, your recent arrival was a great motivator to finish this thesis.

# *Contents*


*Abstract* 2

*Acknowledgements* 4

*List of Figures* 9

*List of Tables* 12

*1. Introduction* 15

*1.1. Familiar and unfamiliar faces* 17
*1.1.1. Face recognition* 18
*1.1.2. Unfamiliar face matching* 21
*1.2. Face matching in applied settings* 24

*2. Face matching expertise* 31

*2.1. Training* 36
*2.1.1. Facial anatomy and photography training* 39
*2.1.2. Short face matching training courses* 40
*2.1.3. Morphological feature comparison* 43
*2.1.4. Feedback training* 49
*2.1.5. Within face variation* 53
*2.1.6. Mentoring* 54
*2.1.7. Training overview* 55
*2.2. Forensic face examiners* 57
*2.2.1. Perceptual skill* 61
*2.2.2. Operational accuracy* 62
*2.2.3. Face examiner expertise* 64
*2.3. Super recognisers* 67
*2.3.1. Group versus individual face matching performance* 71
*2.3.2. Super matchers* 72
*2.3.3. Super recogniser selection* 73
*2.3.4. Are super recognisers experts?* 76
*2.4. Automated facial recognition algorithms* 78
*2.5. Face matching systems* 85
*2.5.1. Group decision making* 86
*2.5.2. Crowd effects in face matching* 88
*2.5.3. Collaborative face matching* 92
*2.5.4. Human computer interactions* 93
*2.5.5. Designing better face matching systems* 97

*3. Thesis overview & research aims* 99

*3.1. Study One – An international survey of face matching training* 100
*3.2. Study Two – The impact of a short training course on face matching behaviour* 101

*3.3. Study Three - Comparing perceptual skill and crowd effects for superior face matchers and face examiners ..... 102*
*3.4. Study Four – Combining human and algorithm expertise ..... 103*
*3.5. Study Five – Operational accuracy of forensic face examiners ..... 103*

*4. Study One – An international survey of face matching training 104*

*4.1. Introduction ..... 104*
*4.2. Methods ..... 105*
*4.2.1. Participants ..... 105*
*4.2.2. Procedure ..... 106*
*4.3. Results and discussion ..... 107*
*4.3.1. Training delivery results ..... 109*
*4.3.2. Training delivery discussion ..... 110*
*4.3.3. Training duration results ..... 111*
*4.3.4. Training duration discussion ..... 113*
*4.3.5. Training topic results ..... 114*
*4.3.6. Training topics discussion ..... 119*
*4.3.7. Evidence-based training strategies results ..... 122*
*4.3.8. Evidence-based training strategies discussion ..... 123*
*4.4. General Discussion ..... 125*

*5. Study two – The impact of a short training course on face-matching behaviour 129*

*5.1. Introduction ..... 129*
*5.2. Method ..... 132*
*5.2.1. Participants ..... 132*
*5.2.2. Materials ..... 132*
*5.2.3. Procedure ..... 133*
*5.3. Results ..... 135*
*5.3.1. Preliminary analysis ..... 135*
*5.3.2. Overall accuracy ..... 137*
*5.3.3. Match and non-match accuracy ..... 140*
*5.3.4. Sensitivity and bias ..... 143*
*5.3.5. High performers and low performers ..... 147*
*5.3.6. Confidence decisions ..... 156*
*5.4. Discussion ..... 165*

*6. Study Three – Comparing perceptual skill and crowd effects for superior face matchers and face examiners 170*

*6.1. Introduction ..... 170*
*6.2. Methods ..... 174*
*6.2.1. Participants ..... 174*
*6.2.2. Materials ..... 174*
*6.2.3. Procedure ..... 174*
*6.3. Results ..... 175*
*6.3.1. Trial A short form and Trial B short form ..... 175*
*6.3.2. Selecting superior face matchers ..... 180*
*6.3.3. Accuracy trial A short form ..... 183*
*6.3.4. Accuracy trial B short form ..... 188*

6.3.5. Sensitivity and response bias of superior face matchers and face examiners ..... 192
6.3.6. Confidence decisions of superior face matchers and face examiners ..... 199
6.3.7. Crowd effects ..... 201
6.4. Discussion ..... 210

7. Study Four – Combining human and algorithm face matching expertise 213

7.1. Introduction ..... 213
7.2. Method ..... 215
7.2.1. Participants ..... 215
7.2.2. Materials ..... 215
7.2.3. Procedure ..... 215
7.3. Results ..... 217
7.3.1. Fusion for human challenging images ..... 217
7.3.2. Fusion for algorithm challenging images ..... 222
7.4. Discussion ..... 229

8. Study Five – Operational accuracy of face examiners 233

8.1. Introduction ..... 233
8.2. Method ..... 236
8.2.1. Participants ..... 236
8.2.2. Materials ..... 236
8.2.3. Procedure ..... 237
8.3. Results ..... 239
8.3.1. Performance ..... 239
8.3.2. Types of errors ..... 241
8.3.3. Consistency of support levels ..... 244
8.3.4. Fusion ..... 246
8.4. Discussion ..... 249

9. General discussion 251

9.1. Summary of findings ..... 251
9.1.1. Training ..... 253
9.1.2. Superior face matchers and crowd effects ..... 258
9.1.3. Operational accuracy of forensic face examiners ..... 261
9.1.4. Combining human and algorithm expertise ..... 263
9.2. Practical recommendations ..... 265
9.2.1. Recommendations for quick decision face matching ..... 265
9.2.2. Recommendations for forensic face matching ..... 267
9.3. Limitations and future research ..... 269
9.4. Conclusion ..... 272

10. References 273

Appendix A – Training survey 292

Appendix B – Face matching training course overview 299

*Appendix C - Face matching trial online consent form* *302*

*Appendix D – Face matching interface* *303*

# *List of Figures*

*Figure 1 – Examples of the same face captured under differing imaging and environmental conditions..... 22*

*Figure 2 – Publications by year on super recognisers and related terms ..... 69*

*Figure 3 – Word cloud of participating countries..... 105*

*Figure 4 – Frequency distributions for durations of examiner and reviewer training..... 112*

*Figure 5 - Example face matching stimuli used in the training evaluation, taken from Towler et al. (2019) (1 - GFMT, 2 - EFCT, 3 images representative of the CWT, 4 - MFMT) ..... 133*

*Figure 6 - Distribution of raw scores on trial A for trainees and controls..... 135*

*Figure 7 - Correlation coefficients for overall accuracy, match and non-match scores between trial A and trial B for the control group (values with black cross are non-significant at the 95% confidence level) ..... 136*

*Figure 8 - Trainee and control accuracy by trial (blue dot represents median group score and red star is the mean group score) ..... 138*

*Figure 9 - Trainee and control accuracy by trial on GFMT, EFCT, CWT and MFMT (blue dot represents median group score and red star is the mean group score)..... 139*

*Figure 10 - Trainee and control match accuracy by trial (blue dot represents median group score and red star is the mean group score)..... 141*

*Figure 11 - Trainee and control non-match accuracy by trial (blue dot represents median group score and red star is the mean group score) ..... 143*

*Figure 12 - Trainee and control sensitivity (A) by trial (blue dot represents median group score and red star is the mean group score)..... 145*

*Figure 13 - Trainee and control bias (b) by trial (blue dot represents median group score and red star is the mean group score) ..... 145*

*Figure 14 - Relationship between bias (b) and sensitivity (A) for trainees and controls on trial A and trial B..... 147*

*Figure 15 - High and low performer A scores by group and trial (blue dot represents median group score and red star is the mean group score)..... 150*

*Figure 16 - High and low performer b scorers by group and trial (blue dot represents median group score and red star is the mean group score)..... 150*

*Figure 17 - High and low performer match scores by group and trial (blue dot represents median group score and red star is the mean group score)..... 154*

*Figure 18 - High and low performer non-match scores by group and trial (blue dot represents median group score and red star is the mean group score)..... 154*

*Figure 19 - Distributions of trainee confidence decisions for trial A and trial B ..... 157*

*Figure 20 - Distributions of control confidence decisions for trial A and trial B ..... 158*

*Figure 21 – Distribution of Spearman's rho values of confidence-accuracy relationship for match pairs by group and trial (blue dot represents median group score and red star is the mean group score) ..... 163*

*Figure 22 – Distribution of Spearman's rho values of confidence-accuracy relationship for non-match pairs by group and trial (blue dot represents median group score and red star is the mean group score) ..... 164*

*Figure 23 – Distribution of scores for all face image pairs from trial A ..... 176*

*Figure 24 – Distribution of scores for 25 hardest matching and 25 hardest non-matching face image pairs from trial A ..... 177*

*Figure 25 – Distribution of scores for all face image pairs from trial B ..... 178*

*Figure 26 – Distribution of scores for 25 hardest matching and 25 hardest non-matching face image pairs from trial B ..... 179*

*Figure 27 - Accuracy for trial A short form by group with control mean, 1 SD, 1.5 SD and 2 SD cut offs ..... 181*

*Figure 28 – Match and non-match accuracy of superior face matchers on Trial A short form ..... 183*

*Figure 29 – Accuracy on trial A short form by group (blue circle represents median group score and red star is the mean group score) ..... 185*

*Figure 30 – Scatterplot of match and non-match accuracy by group on trial A short form ..... 186*

*Figure 31 – Accuracy on trial B short form by group (blue circle represents median group score and red star is the mean group score) ..... 189*

*Figure 32 – Scatterplot of match and non-match accuracy by group on trial B short form ..... 190*

*Figure 33 – Sensitivity (A) by trial for individual SMs and controls (errors bars represent one standard deviation from the control mean) ..... 194*

*Figure 34 – Response bias (b) by trial for individual SMs and controls (errors bars represent one standard deviation from the control mean) ..... 194*

*Figure 35 – Sensitivity (A) by trial for individual FEs and controls (errors bars represent one standard deviation from the control mean) ..... 195*

*Figure 36 – Response bias (b) by trial for individual FEs and controls (errors bars represent one standard deviation from the control mean) ..... 195*

*Figure 37 – Boxplots of proportion of errors rate by confidence for each group ..... 200*

*Figure 38 – Accuracy on trial B short form for controls, SMs and SM crowds (blue circle represents median group score and red star is the mean group score) ..... 204*

*Figure 39 – Scatterplot of match and non-match accuracy for controls, SMs and SM crowds on trial B short form ..... 204*

*Figure 40 – Accuracy on trial B short form for controls, FEs and FE crowds (blue circle represents median group score and red star is the mean group score) ..... 206*

*Figure 41 – Scatterplot of match and non-match accuracy for controls, FEs and FE crowds on trial B short form ..... 207*

*Figure 42 – Distribution of individual SM confidence decisions for trial B short form..... 209*

*Figure 43 – Distribution of SM pair confidence decisions for trial B short form..... 209*

*Figure 44 – AUC scores by group for human challenging face pairs (grey dashed line represents algorithm score)..... 218*

*Figure 45 – Scatterplot of human AUC scores and fused AUC scores for human challenging images (grey dashed lines represent algorithm score) ..... 219*

*Figure 46 – AUC scores for the human group separated by performance on human challenging face pairs (grey dashed line represents algorithm score) ..... 220*

*Figure 47 – Scatterplot of human AUC scores and scores difference after fusion for human challenging images (grey dashed line represent algorithm score) ..... 222*

*Figure 48 – Frequency distribution of algorithm scores for matching and non-matching pairs on trial A ..... 223*

*Figure 49 – AUC scores by group for algorithm challenging face pairs (grey dashed line represents algorithm score)..... 225*

*Figure 50 – Scatterplot of human AUC scores and fused AUC scores for algorithm challenging images (grey dashed lines represent algorithm score) ..... 225*

*Figure 51 – AUC scores separated by performance on algorithm challenging face pairs (grey dashed line represents algorithm score) ..... 227*

*Figure 52 – Scatterplot of human AUC scores and scores difference after fusion for algorithm challenging images (grey dashed line represent algorithm score) ..... 228*

*Figure 53 - AUC scores for ENFSI test by group (blue circle represents median group score and red star is the mean group score..... 240*

*Figure 54 – Distributions of proportions of error rate by support level for controls, individual examiners and examiner teams ..... 242*

*Figure 55 – Distributions of proportion of no support decisions by group for ENFSI test 244*

*Figure 56 – Median correlation coefficients of decision rating agreement for all human participants by group ..... 246*

*Figure 57 – AUC scores by group for the ENFSI test images (grey dashed line represents algorithm score)..... 248*

*Figure 58 – Recommendations for quick decision face-matching scenarios ..... 266*

*Figure 59 – Recommendations for forensic face matching scenarios..... 267*

# *List of Tables*

*Table 1 – List of FISWG feature components ..... 45*
*Table 2 – FISWG characteristic descriptors of the nose ..... 46*
*Table 3 – Accuracy improvements from face matching training ..... 56*
*Table 4 – Overview of training type, delivery method and duration by agency ..... 108*
*Table 5 – Source of training by agency ..... 109*
*Table 6 - Delivery methods for reviewer and examiner training ..... 109*
*Table 7 – Duration of reviewer and examiner training ..... 112*
*Table 8 – Training topics covered by examiner and reviewer training ..... 115*
*Table 9 – Anatomy training subtopics covered by reviewer and examiner training ..... 116*
*Table 10 – Image science subtopics covered by reviewer and examiner training ..... 117*
*Table 11 – Image processing subtopics covered by reviewer and examiner training ..... 117*
*Table 12 – Comparison method subtopics covered by reviewer and examiner training . 118*
*Table 13 – Evidence-based approaches to improving face matching accuracy ..... 122*
*Table 14 – Research-based approaches used in reviewer and examiner training ..... 123*
*Table 15 - Summary statistics for trainee and control overall accuracy ..... 137*
*Table 16 - Summary statistics for trainee and control match accuracy ..... 140*
*Table 17 - Summary statistics for trainee and control non-match accuracy ..... 142*
*Table 18 - Summary statistics for trainee and control sensitivity (A) and bias (b) ..... 144*
*Table 19 - Summary statistics for low trainee and high trainee sensitivity (A) and bias (b) ..... 148*
*Table 20 - Summary statistics for low control and high control sensitivity (A) and bias (b) ..... 149*
*Table 21 - Summary statistics for low performing and high performing trainee match accuracy and non-match accuracy ..... 152*
*Table 22 - Summary statistics for low performing and high performing control match accuracy and non-match accuracy ..... 153*
*Table 23 – Summary statistics of confidence decisions differences between trial B and trial A for trainee group ..... 159*
*Table 24 – Summary statistics of confidence decisions differences between trial B and trial A for trainee group with outlier removed ..... 160*
*Table 25 – Summary statistics of confidence decisions differences between trial B and trial A for control group with outlier removed ..... 161*
*Table 26 - Summary statistics of Spearman's rho values of confidence-accuracy relationship for match pairs ..... 162*

*Table 27 - Summary statistics of Spearman's rho values of confidence-accuracy relationship for non-match pairs* ..... *162*

*Table 28 - Summary statistics of item difficulty for Trial A long form and Trial A short form* ..... *177*

*Table 29 – Summary statistics of item difficulty for Trial B long form and Trial B short form* ..... *179*

*Table 30 – Summary statistics of control and selection pool accuracy on trial A short form* ..... *181*

*Table 31 – Summary statistics of overall, match and non-match accuracy for Trial A short form by group* ..... *184*

*Table 32 – Individual case analyses comparing accuracy of superior face matchers with mean control accuracy on trial A short form* ..... *187*

*Table 33 – Individual case analyses comparing accuracy of superior face matchers with mean control accuracy on trial A short form* ..... *187*

*Table 34 – Summary statistics of overall, match and non-match accuracy for Trial B short form by group* ..... *188*

*Table 35 – Individual case analyses comparing accuracy of superior face matchers with mean control accuracy on trial B short form* ..... *191*

*Table 36 – Individual case analyses comparing accuracy of face examiners with mean control accuracy on trial B short form* ..... *192*

*Table 37 – Summary statistics of A and b for Trial A short form and Trial B short form by group* ..... *193*

*Table 38 – Individual difference analyses comparing sensitivity (A) of superior face matchers between trials with mean control A* ..... *197*

*Table 39 – Individual difference analyses comparing response bias (b) of superior face matchers between trials with mean control b* ..... *197*

*Table 40 – Individual difference analyses comparing sensitivity (A) of face examiners between trials with mean control A* ..... *198*

*Table 41 – Individual difference analyses comparing response bias (b) of face examiners between trials with mean control b* ..... *198*

*Table 42 – Summary statistics of accuracy for SM and FE crowds on trial B short form* *203*

*Table 43 – Individual case analyses comparing accuracy of SM crowds with mean control accuracy on trial B short form* ..... *205*

*Table 44 – Individual case analyses comparing accuracy of FE crowds with mean control accuracy on trial B short form* ..... *208*

*Table 45 – Summary statistics of AUC scores for the algorithm, humans and fusion results on human challenging images from trial A* ..... *217*

*Table 46 – Summary statistics of human and fused AUC scores split by performance on human challenging images from trial A* ..... *220*

*Table 47 – Summary statistics of AUC scores for the algorithm, humans and fusion on algorithm-challenging images from trial A* ..... *223*

*Table 48 – Summary statistics of human and fused AUC scores split by performance on algorithm challenging images from trial A.....227*

*Table 49 – Summary statistics of AUC scores by group for ENFSI test images.....239*

*Table 50 – Summary statistics for correlation coefficients of decision rating agreement by group .....245*

*Table 51 – Summary statistics of AUC scores for the algorithm, unfused and fused groups on the ENFSI test images.....247*

# *1. Introduction*

Face matching is widely used as a means of identification in high stakes, security critical settings, such as law enforcement, forensics, defence and border security. Examples include verifying a passport holders' identity at the border, comparing images of a suspect to CCTV of an offender or identifying persons of interest in public spaces using automated facial recognition. In applied settings the outcome of a face-matching decision could have potentially life changing consequences, such as a person being denied entry at the border, a suspect's arrest in a police investigation or a defendant being convicted and sentenced to imprisonment. Given the societal implications of a misidentification, facial identification has, justifiably, received heightened scrutiny. Recent media articles have been highly critical of the uptake of automated facial recognition systems by governments and police forces, raising concerns of discrimination and stifling of free speech (Booth, 2020). Another article highlighted the fallibility of police super recognisers and the limited understanding of their abilities (Moshakis, 2018) and there has also been a recent challenge to the scientific basis of facial image comparison techniques used by forensic experts (Gabrielson, 2019). Each of these articles discusses applied face matching but in each case the source of the face matching decision is different; an algorithm trained to match faces and compute similarity scores, a super recogniser who is believed to be naturally proficient at matching faces and a forensic examiner trained to carry out detailed face matching examinations. It is essential, therefore, to understand how each of these three different face-matching sources operates and to validate the veracity of any associated claims of expertise in face matching, such that the strengths and limitations of each source are understood.

In applied settings face-matching tasks are carried out by face-matching *systems*, where individual human operators and computer programs each form a component of the system and provide a source of face-matching expertise (Towler, Kemp, et al., 2017). It is, therefore, necessary to understand how different components within the system interact and the impact of these interactions on face-matching decisions. For example, multiple forensic examiners may be involved in a face-matching examination and contribute to the end result (Moreton, 2021). Even where automated facial recognition systems are used to search a facial image database, the results must be verified by a human operator (White, Dunn, et al., 2015). The combined involvement and interaction of multiple human operators and computer algorithms can have a significant bearing on the overall accuracy of the system. If the individual components are properly integrated this will lead to overall gains in system accuracy, conversely if a system is poorly designed and components are not well integrated there may be deleterious effects on system accuracy (White, Dunn, et al., 2015). In order to optimise the overall accuracy of face-matching systems it is necessary to understand the expertise and interaction of different face-matching components at both the individual and system level.

The next section provides an introduction to the cognitive processing of familiar and unfamiliar faces, followed by an overview of face matching in applied settings. Chapter 2 then delivers a critical review of the face-matching expertise literature, looking specifically at four sources of expertise in applied face-matching systems: professional training; forensic face examiners; super recognisers; and automated facial recognition algorithms. This critical review of the literature is followed by a series of empirical studies investigating the different sources of face-matching expertise used in applied face-matching systems.

## *1.1. Familiar and unfamiliar faces*

There are a wide range of face identification tasks in applied settings, from matching a live person to an image at a border, to visual searches of faces in a crowd or trawling databases of known faces to find a match to an unidentified image (Moreton et al., 2019). Applied face identification tasks can be broadly separated into the following three categories:

**Unfamiliar face matching** - the direct comparison of two or more images of a face, or an image to a live subject, to determine whether they depict/are the same individual. The decision is based upon perception and comparison, it is not a task that utilises memory. This is also referred to as facial comparison or facial examination within a forensic context (Facial Identification Scientific Working Group, 2010).

**Familiar facial recognition** - The identification of a live subject, or a subject depicted in an image, that the observer has met or seen previously and is familiar with. Generally familiar face recognition is effortless, instantaneous and possible even from low quality imagery. The observers innate ability at recognition (Russell et al., 2009) and how familiar they are with the subject (Bruce & Young, 1986) are factors that impact upon the reliability of familiar face recognition.

**Unfamiliar facial recognition** - The identification of a face that the observer has briefly been exposed to. Typically, there will only have been a short window of exposure and the observer will not be considered familiar with the subject. Examples of such tasks in applied settings include recalling a face seen at an event (eye-witness identification) or recognising a subject in CCTV from an image viewed earlier in the day.

A clear distinction in the perceptual processes underpinning these different types of face identification is whether the faces are familiar or unfamiliar to the observer. There are

fundamental differences in the processes and also the difficultly of familiar and unfamiliar face identification (Megreya & Burton, 2006). Recognising faces we know tends to be accomplished quickly and with a high degree of accuracy even from poor quality images, whereas matching faces we do not know can be very challenging (Bruce et al., 2001). Despite these differences it is commonplace for tasks such as familiar face recognition and unfamiliar face matching to be referred to under the umbrella terms 'facial identification' and 'facial recognition'. From an applied perspective, differences between the identification of familiar and unfamiliar faces are poorly understood and frequently confused. Before moving onto applied face matching, the processes that underlie how we identify familiar faces will be briefly introduced, to establish that whilst familiar and unfamiliar face identification processes are related there are fundamental differences between them.

### 1.1.1. Face recognition

In 1986 Bruce & Young published a model describing how humans encode and recognise familiar faces. The Bruce and Young model postulates that each familiar individual is stored in memory within theoretical Facial Recognition Units (FRUs). The information relating to that identity is encoded in different types of codes. Structural, identity specific semantic and name codes play the predominant role in recognising familiar faces. Name codes store the known name associated with a face, whereas identity specific semantic codes contain non-face information related to the context of how an individual is known or where they were met, such as their job or a specific location where they are commonly encountered. Identity specific semantic codes are suggested as the reason why it is easier to recognise a familiar individual when they are observed in a relevant context and harder when the individual is seen out of context. Structural codes store the relevant facial information used to recognise an individual. The alternative face space model of human facial recognition proposed by Valentine, (1991) conceptualises familiar face memory as a multi-dimensional space where

new encounters of a familiar face will be encoded at a distance or vector from the typical representation of that face. The distance will increase the more that a new encounter varies from the typical face, with the typical face defined by experience and exposure to different faces. Familiar faces stored in memory can be updated with new information from novel encounters, such as from a different perspective, a new expression or different lighting conditions, meaning the more familiar we are with a face the easier it is to recognise that individual despite the conditions in which they are encountered.

For over 30 years it has been theorised that the recognition of familiar faces is based upon configural face information, including first-order relations between features, holistic processing and the spacing between features, rather than the processing of specific features themselves (Maurer et al., 2002). The importance of configural face processing is derived from the inversion effect, first demonstrated by (Yin, 1969), where face identification accuracy is impaired when faces are inverted. Research has also demonstrated that individual facial features have a limited role in familiar recognition. For example, reliable recognition can be achieved from low quality images where specific features cannot be resolved and only low frequency facial information is available (i.e. the overall shape and texture of the face) (Costen et al., 1996). This ability to identify familiar faces using only low frequency information provides a possible explanation for the greater ease with which we can recognise known faces from low quality images, in contrast to the comparison of unfamiliar faces we do not know (Bindemann et al., 2013). However, the configural face processing account has been criticised by Burton et al. (2015) as ill-defined in the literature and poorly evidenced by empirical research, particularly given that extreme geometric distortions of configural face information do not appear to impair recognition accuracy (Hole et al., 2002). Instead, Burton et al. (2015) believe that familiarity with a specific face is the

major contributor to the enhanced accuracy with which we identify people we know, rather than solely a reliance on configural processing.

Since the development of these early models of face recognition, research into how humans encode and identify faces has progressed somewhat slowly, due in part to studies using overly artificial stimuli and conflating familiar and unfamiliar faces, which are perceived in qualitatively different ways (Burton, 2013). Past research in familiar face recognition has also been overly concerned with discriminability, or the ability to tell different faces apart (Jenkins et al., 2011). More recently, researchers have focussed on the importance of *within-face variability* as an explanation for why it is so much harder to identify unfamiliar faces compared to familiar faces (Andrews et al., 2015). Within-face variability is believed to be idiosyncratic and must be learned for each new identity (i.e. this variability does not generalise to new faces) (Burton et al., 2016; Kramer et al., 2018). This theory is revisited in more detail in Chapter 2.

Recognising unfamiliar faces that we have been exposed to briefly is believed to be a process distinct from recognising faces with which we have a high degree of familiarity (Bruce et al., 2001; Megreya & Burton, 2006; Young & Burton, 2018). In contrast to well-learned familiar faces, if a face is only encountered briefly, or from a single image, a robust memory of that person cannot be created due to the limited conditions inherent in a brief exposure. Instead, Bruce & Young (1986) stated that a simpler pictorial code will be generated without the variation of a more established, robust FRU. If only a basic pictorial code is created it is less likely that an individual will be recognised in a new encounter, such as due to changes in expression (Bruce, 1982), viewpoint (O'Toole et al., 1998) or different lighting conditions (Etchells et al., 2016).

The Cambridge Face Memory Test (CFMT) is an example of an unfamiliar facial recognition task that has been extensively used by researchers in face perception (Duchaine & Nakayama, 2006). Participants briefly observe a face and are then tasked with selecting the matching face from an array without the original stimulus present. The tasks get progressively more challenging with variations in pose and decreasing image quality. The CFMT was initially used to diagnose developmental prosopagnosia, or 'face blindness', where individuals are unable to process or recognise faces. More recently the test has been used to identify individuals with higher than usual facial recognition ability, termed super recognisers (Russell et al., 2009).

### 1.1.2. Unfamiliar face matching

Unfamiliar face matching is not recognition *per se* as there is no reliance on memory. Studies instead place unfamiliar face matching as being more akin to non-face specific tasks such as object and pattern matching (Megreya & Burton, 2006). In contrast to the high accuracy of familiar facial identification, performance at identifying unfamiliar faces is relatively poor (Bruce et al., 2001). Because the observer will have not seen the faces before they will not know the extent of within-face variability that exists for those faces and the identification decision must be based solely on information provided in an image or single encounter. This is in contrast to familiar faces where idiosyncratic variability is learnt through repeated exposure to individuals in different environments (Young & Burton, 2018). Even when unfamiliar faces are photographed on the same day under controlled conditions observers make an incorrect matching decision on average one fifth of the time (Burton et al., 2010).

Performance in unfamiliar face matching further decreases when images are uncontrolled, e.g. captured under different imaging conditions (Dowsett & Burton, 2015) or by different

devices (Burton et al., 2001). Uncontrolled imaging conditions introduce variation into the appearance of an individual face, as shown in the images in Figure 1 all of which are of the same person.


*Figure 1 – Examples of the same face captured under differing imaging and environmental conditions*

Image quality factors that have been specifically shown to impact on unfamiliar face matching include the spatial resolution of digital images (Bindemann et al., 2013), where decreasing the number of pixels within the image reduces the level of visible facial detail, compression of imagery during recording and/or transmission causing a loss of detail, and introducing erroneous image artefacts (Keval & Sasse, 2008) and lighting conditions at time

of capture (Tsifouti, 2016). Also, age differences (Megreya et al., 2013), differences in ethnicity between the observer and the subject in the image (Megreya et al., 2011) and variation in expression and pose (Jenkins et al., 2011) are all detrimental to unfamiliar face-matching ability. Even the wearing of glasses has been shown to impair matching accuracy (Kramer & Ritchie, 2016). These detrimental factors are additive and when present in combination they will further decrease the likelihood of a correct matching decision being made. This is particularly important in applied settings where decisions are made based on real-world, uncontrolled imagery such as CCTV video or operational surveillance images, as multiple factors may each introduce distortions or artefacts into the imagery (Seckiner et al., 2018). In operational settings it can only be assumed that the error rates reported in the literature will be greater due to the uncontrolled nature of the imagery being compared. In addition to the issues around image quality discussed above, studies have also shown poor consistency on unfamiliar face-matching tasks, where the observers may not reach the same decision on image pairs compared on different days (Bindemann et al., 2012). Individuals can also be easily swayed into making erroneous matching decisions by contextual information (Sauerland et al., 2016). There is a wide range in face-matching ability between different individuals, with some individuals achieving perfect accuracy whereas others perform close to chance on the same test (Burton et al., 2010). Recently, there has been an increased focus on individual differences in both familiar face recognition and unfamiliar face matching (Lander et al., 2018; Wilmer, 2017).

The issues of generally poor performance and the extent of individual differences in face matching are further exacerbated by individuals having poor insight into their own face-matching ability, conflating the task with the more accurate process of familiar facial recognition (Bindemann et al., 2014). Zhou & Jenkins (2020) have recently found evidence for a Dunning-Kruger effect in people's perception of their face-matching ability, where

individuals with low face-matching ability over-estimated their ability and high performers believed others to have greater ability than they actually possess. This lack of insight into face-matching ability is of particular significance for operational face-matching personnel, where errors can have a significant impact on people's lives. As a result, the consideration of who conducts face matching in applied settings and why they are selected to do so is of great importance.

## *1.2. Face matching in applied settings*

There is a long and established history of using facial images as a means of identification in applied settings. Facial images were a mandatory requirement for UK passports from 1915, introduced to safeguard national security amid fears that foreign spies could too easily pass through the border during the First World War[1]. The use of 'mug shot' images by policing goes back even further, beginning in the 1840s. The UK Home Office then released a standard for the taking of prisoner photographs in 1890[2]. Today the use of facial images as a means of identity verification is ubiquitous and increasingly becoming automated. Major UK airports have e-gates to allow automatic verification of travellers at the border, the UK police national database enables the searching of over 19 million custody facial images via an automated facial recognition algorithm and it is admissible for forensic experts to present face matching evidence in the courtrooms. For most applied uses of face

[1] https://www.bbc.co.uk/news/magazine-30988833

[2] https://blog.scienceandmediamuseum.org.uk/a-z-of-national-photography-collection-m-is-for-mugshots/

matching (even in cases where an automated algorithm is used) a human is required to act as the final arbiter of the match decision.

Despite the widespread use of facial images to verify identity, research has consistently shown that untrained persons are, on average, surprisingly poor at comparing faces of people they do not know (Burton et al., 2010; Dowsett & Burton, 2015; Fysh & Bindemann, 2017a; Kemp et al., 1997). Even when the facial images being compared are taken in controlled conditions on the same day, untrained participants are mistaken on average 20% of the time (Burton et al., 2010). Accuracy between individuals is also highly variable with performance ranging from perfect to almost chance on the same images. Accuracy further decreases when images are of low quality or taken in uncontrolled conditions such as from CCTV systems (Burton et al., 1999), as is often the case operationally. This variability in human accuracy is a significant issue for operational settings where critical decisions may be based on the comparison of facial imagery, often of low quality.

The fact that face-matching ability varies significantly between different individuals makes a challenging dilemma for how face matching can be used reliably in applied settings. It seems logical that through repeated experience of matching faces individuals should develop some perceptual expertise in the task. Perceptual learning in sensory tasks can be developed through repeated practice at a task (Hussain et al., 2009a) and repeating a perceptual task improves an observer's ability to discriminate between different stimuli. However, these benefits are specific to the stimuli used in the practice stage (Hussain et al., 2009b) unless large and diverse training sets are used (Hussain, McGraw, et al., 2012). Feedback also appears to be a critical component for inducing perceptual learning, providing a mechanism for individuals to learn from their mistakes (White, Kemp, Jenkins, & Burton, 2014), but results have differed as to the specific benefits of feedback for improving accuracy in relation to face matching (Alenezi & Bindemann, 2013). In applied

settings like policing, operational staff will be exposed to high volumes of faces that may contain diverse stimuli, depending upon their role (e.g. CCTV, passports, surveillance imagery, social media). Therefore the quantity and diversity of face matching decision made by operational staff may provide some basis for perceptual learning, but the ground truth of operational images is unknown and staff seldom receive feedback on their accuracy or the correctness of their decisions. Although the idea of work-based experience in face matching providing a route to perceptual expertise seems tangible, the small number of studies conducted in applied settings show quite the opposite.

Kemp et al. (1997) published one of the earliest studies demonstrating the fallibility of face matching in an applied setting. Six shop cashiers were presented with photo ID credit cards to verify the identity of shoppers. Images were captured less than 6 weeks prior to the experiment to mitigate the impact of age-related changes. Even in these optimised conditions cashiers only made the correct decision on average 67.4% of the time and only 36.3% of decisions were correct when the card did not depict the shopper (correct rejections) (Kemp et al., 1997). As the cashiers in this study did not have any specific experience or training in face matching, it could be argued that this was a contributory factor to their poor performance. However, further studies testing operational personnel with training and/or experience in face-matching tasks have shown similarly poor levels of performance. For example, when a group of police officers were tasked with matching faces from analogue CCTV footage to high quality face images their performance was no better than undergraduate psychology student controls (Burton et al., 1999). White, Kemp, Jenkins, Matheson, et al. (2014) evaluated the performance of Australian passport officers in both a live subject to image matching task, and an image to image matching task. In both tests passport officer performance was comparable to controls, and surprisingly the study found no relationship between duration of employment and face-matching accuracy on

either trial. White et al. propose that this observation is not limited to the participants of this study but may be common to many applied settings where operational staff are making face matching decisions.

Wirth & Carbon (2017) explored the relationship between employment duration and accuracy, testing 96 German police officers working in border protection on a facial image matching task. Performance of the officers was compared to 48 student controls. The police officers were observed to perform significantly better than students (though still making a high proportion of errors, particularly on mismatch trials). Wirth & Carbon conducted further analysis of the police officer group and found, surprisingly, that it was only the officers with shorter durations of employment who had recently completed training that outperformed controls, whereas those with longer terms of employment did not perform significantly better. The authors hypothesise that having recently completed training contributed to the enhanced performance of the police officers with shorter employment, unfortunately the study does not elaborate on how these police officers are trained or whether they are pre-selected for the role.

Recently, White et al. (2021) completed a meta-analysis of 12 published studies evaluating the performance of face-matching professionals in 25 different experiments. The meta-analysis included three different groups of face-matching professionals: facial reviewers; face examiners; and police super recognisers. The first two groups were defined using guidance documents produced by the international Facial Identification Scientific Working Group (FISWG). Facial reviewers are a diverse group of trained professionals, including bank tellers, police officers, border control officers and passport issuance officers, who conduct face matching in high-throughput environments, often to provide investigative and operational leads. Facial reviewers may also work with automated facial recognition systems (Facial Identification Scientific Working Group, 2019c). Face examiners are

specialised face matching professionals, who commonly work in small teams within police departments, forensic service providers and government agencies. Face examiners conduct face matching by rigorous morphological analysis, comparison and evaluation of images, often in a forensic setting (Facial Identification Scientific Working Group, 2019b). The third group, police super recognisers, consisted of face-matching professionals selected based on their superior innate ability on various face identification tasks. However, the specific mechanisms by which police super recognisers are recruited and selected in different organisations are not well understood in the literature (White et al., 2021).

The meta-analysis showed mixed results between the three professional groups. Concerningly, in 12 out of 18 experiments facial reviewers showed no improvement in face matching accuracy over untrained controls and *lower* performance in half of studies. These findings suggest that the professional experience of the majority of facial reviewers gave no advantage in face matching accuracy, despite this task being a primary part of their role. Although agencies provide training for facial reviewers it is often only for short durations (Heyer, 2013). As will be demonstrated in later chapters, there is limited validation of the effectiveness of short training courses, and what data does exist shows little impact on improving face-matching accuracy. Another possible cause of the varied and generally poor performance of the facial reviewer group is the heterogenous makeup of the group. Facial reviewers are employed in a wide range of jobs carrying out very different tasks. From a survey of facial reviewers in Australia, Heyer, (2013) found a wide diversity of educational backgrounds, training, experience and attitudes towards face matching.

In contrast with findings that show that training and professional experience offer no advantage for many facial reviewer groups (Burton et al., 1999; White, Kemp, Jenkins, Matheson, et al., 2014; Wirth & Carbon, 2017), professional face examiners were shown to consistently outperform control groups in all seven tests reviewed in the meta-analysis

(Norell et al., 2015; Phillips et al., 2018; Towler, White, et al., 2017; White, Dunn, et al., 2015; White, Kemp, Jenkins, Matheson, et al., 2014). Face examiners often provide expert evidence on facial image comparisons in the courtroom, and undergo lengthy training and mentoring in face matching, employing detailed and rigorous methods to compare the images (Houlton & Steyn, 2018). As a group, examiners have been shown to perform highly relative to untrained controls in both quick decision face matching tests (White, Phillips, et al., 2015) and when allowed to use their own tools and methods (Norell et al., 2015; Phillips et al., 2018). Examiners have also been demonstrated to be more cautious in making comparison decisions particularly from low quality images (Norell et al., 2015), possibly because they are more aware of the likely impact that image quality can have, and mindful of the ramifications of an incorrect decision in an applied setting.

It is important to note that forensic face examiners are not a homogenous group and significant variation in accuracy exists between different examiners on the same test. There has also been heavy criticism of certain working practices used by forensic face examiners, such as the technique of measuring facial proportions (Kleinberg & Vanezis, 2007; Moreton & Morley, 2011) and overlaying or superimposing facial images (Strathie et al., 2012; Strathie & McNeill, 2016). For detailed reviews see Edmond et al. (2009); Mallett & Evison (2013) and McNeill et al. (2015). Despite past criticisms, the accuracy of examiners at group level has been consistently demonstrated (Phillips et al., 2018; White, Phillips, et al., 2015). This suggests that further scrutiny of the work of forensic face examiners is required to establish what it is about this group that gives them their enhanced performance, whether this be training and mentoring, professional expertise, innate perceptual skill or a combination of multiple factors.

Police super recognisers also outperformed untrained control participants in all of the face matching experiments reviewed by White et al. (2021) (see Davis et al., 2016 and

Robertson et al., 2016). Super recognisers are individuals suggested to have an enhanced ability in face identification without any training or professional experience. This enhanced ability can relate to both recognition and matching tasks (though this does not always present in the same manner across different individuals (e.g. Davis et al., 2016)). Although the complexities of defining what exactly qualifies someone as a super recogniser are only now being addressed (Noyes et al., 2017; Noyes & O'Toole, 2017; Ramon et al., 2019a), it is clear that based on the wide range in innate face matching ability, organisations could benefit from factoring this variance into the selection and deployment of face-matching personnel. Alongside training and expertise, selection based on face-matching ability would appear to be another obvious solution to the face-matching problem.

Automated facial recognition algorithms are now widely used in applied settings for a variety of tasks, including one-to-one verification (e.g. at passport control), searching large databases (also known as one-to-many identification) and grouping sets of face images (sometimes referred to as clustering) (Noyes & Hill, 2021). Whilst the media has generally been sceptical and often highly critical of the accuracy of automated facial recognition algorithms, huge gains in accuracy have been made in recent years (Grother et al., 2019a). Only a small number of studies have directly compared human and algorithm face matching accuracy on the same stimuli, but those that have, have shown a clear upward trajectory for algorithm performance. Whereas in 2007 only a small number of algorithms could rival the accuracy of untrained human participants matching controlled, frontal, full-face images (O'Toole et al., 2008), a little over 10 years later state of the art deep convolution neural networks (DCNNs) have shown comparable levels of accuracy to some face examiners and super recognisers on a challenging face matching task (Phillips et al., 2018).

# *2. Face-matching expertise*

In order to improve accuracy in face matching tasks it is necessary to comprehend the nature of expertise that underpins enhanced face-matching performance. Edmond et al. (2017) suggested that a thorough understanding of expertise can lead to the design of more effective training regimes, assist in identifying candidates more likely to develop expertise and inform the design of working environments that improve performance. In the context of applied face matching these points can be interpreted as training procedures that increase the likelihood of correct face matching decisions, identifying individuals with superior face matching ability or the potential to develop such abilities and building face-matching systems that result in overall improved performance, as advocated by Towler, Kemp, et al. (2017). Expertise has been studied across a wide range of domains from the expert strategies of chess masters (Gobet & Charness, 2006) to the navigational skills of London taxi drivers (Maguire et al., 2000). Whilst there is no single accepted definition of what an expert is in the literature, it is broadly understood by cognitive scientists to mean somebody who shows consistently superior performance in a specific task, acquired by repeated practice and experience (Skovholt et al., 2016). Expert performance should also be highly reproducible and show a large, reliable difference to the performance of novices (Ericsson & Lehmann, 1996). Expertise comprises the mechanisms that underlie an expert's superior performance (Edmond et al., 2017).

The knowledge and experience that accompanies expertise is often highly specific to a particular domain and does not generalise to novel domains (Baer, 2015). Despite the domain specificity of experts, there are common markers of expertise identified in the literature. For example, experts have vastly superior memory for domain specific content

compared to novices, even when that information is presented briefly (Ericsson & Lehmann, 1996). This phenomenon is referred to as the 'skilled memory theory', whereby experts can rapidly encode and retrieve domain specific content due to their prior knowledge and greater organisational structure of memory (Ericsson & Staszewski, 1989). This allows experts to draw upon a wider range of problem-solving strategies compared to novices, identifying the solution as they comprehend the problem (Skovholt et al., 2016). Experts are also able to understand domain specific problems at a deeper level than novices, using meaningful features to organise and solve problems, and identify stimuli at more specific and subordinate levels (Tanaka et al., 2005).

The ease with which humans can recognise familiar faces, even from a fleeting glance with high levels of accuracy and automaticity, is evidence that we possess expertise in the processing of faces. How this expertise arises is a subject much debated in the literature. There are broadly two camps of thought on the topic, one being that human expertise in face perception arises through the extensive experience and exposure to faces, the 'expertise hypothesis' and the other being the 'domain specificity' hypothesis, which postulates that human's possess cognitive process and neural substrates that are face specific (McKone et al., 2007). Evidence for the 'domain specificity' argument is largely derived from studies measuring neurological responses to faces. Event-related potentials (ERPs) are measured responses of the brain to specific sensory, cognitive or motor stimuli. The N170 ERP component has been observed to increase in amplitude when observers are shown faces (including non-human faces) but not when observers are shown other types of objects, such as cars and birds (Carmel & Bentin, 2002). Increases in N170 have been observed to occur strongly in specific parts of the right hemisphere of the brain within the fusiform gyrus, dubbed the fusiform face area (FFA). The FFA is understood to contribute towards both face detection processes, face individualisation and face

categorisation (Ghuman et al., 2014). Researchers thus argue that, because of the specificity of the N170 ERP component and the FFA, humans have evolved neural substrates that are specific to faces (Young & Burton, 2018). However, research has indicated that activation of the FFA *does* occur for other stimuli when carrying out within-category identification. Gauthier et al. (2000) found FFA activations to occur when bird and car experts where tasked with categorising different images of birds and cars. Therefore, it is task specificity rather than solely face specificity that defines the function of the FFA.

The 'expertise hypothesis' argues that it is through continuous exposure to faces that humans develop face-specific expertise, explaining why we can detect faces and individualise familiar faces with high levels of accuracy and autonomy (Diamond & Carey, 1986). The perceptual learning route for face expertise has also been used to explain the other race effect (ORE), where recognition ability is poorer with faces that are not the same ethnicity as the observer (Lucas et al., 2011). ORE has also been observed to decrease through exposure to other ethnicity faces (Meissner & Brigham, 2001), which further supports a perceptual learning mechanism for face expertise.

Whilst the literature has yet to reach a consensus on the source of human face expertise, Young & Burton, (2018) argue that many studies overlook a key distinction in deciding whether we are 'naturally' face experts, which is whether faces are familiar or unfamiliar to the observer. For familiar face recognition, the majority of people demonstrate expertise in the task. In general, familiar face recognition is both highly accurate and autonomous when faces are well learned. However, based on years of empirical research demonstrating the difficulty of unfamiliar face matching (see Section 1.1.2), the argument for naturally occurring expertise in unfamiliar face matching is not supported. On average, people are much poorer at matching unfamiliar faces and whilst the task can be performed very quickly,

research has shown that accuracy can be impaired by very short response times (<100ms), which is not the case for familiar faces (Yan et al., 2017).

Young & Burton, (2018) theorise that humans are experts in familiar faces because the variability of the face has been learned through repeated exposure, which allows us to recognise a familiar face accurately in novel encounters. However, this variability is idiosyncratic to a particular face and does not generalise to other faces. The idiosyncratic nature of face variability has been demonstrated in computational modelling of face variability, using dimensionality reduction and machine learning techniques (Burton et al., 2016; Kramer et al., 2018). As a result of this idiosyncrasy our expertise with familiar faces is identity specific and cannot be applied to new, unfamiliar faces. This provides a feasible explanation for the substantially lower levels of accuracy for unfamiliar face memory and matching tasks in the literature (Young & Burton, 2017). But the literature is yet to fully understand why some individuals show consistently high levels of accuracy in unfamiliar face matching whereas others do not.

Understanding that we are not all 'natural' experts at unfamiliar face matching is profoundly important not just for directing future research but also in applied settings where face matching is used in high risk environments. Towler et al. (2021) present a case for two routes to expertise in unfamiliar face matching. The first is derived from the core face processing system and is largely, if not entirely, untrainable based on research findings to date. This route can be considered a person's natural ability, which for some, such as super recognisers, involves performance at levels that could be considered expert, but for most people does not. The second route is the slow, feature based face matching used by trained face examiners. This featural-based approach is highly specialised and distinct from how people naturally match faces. These processes also appear to be derived from intentional training rather natural, perceptual learning of faces. A third route to face-matching expertise

that should also be considered is face matching by computer algorithms. State of the art facial recognition algorithms have improved substantially since 2014 with the introduction of deep convolutional neural networks. State of the art algorithms have been demonstrated to perform at comparable levels to both face examiners and super recognisers on a challenging face matching task (Phillips et al., 2018).

Evaluating the different routes to face matching expertise, i.e. training strategies, face examiners, super recognisers and computer algorithms, will help improve understanding of how face matching accuracy can be maximised in applied settings. However, experts can and do make errors. Often these errors can be directly caused by expertise itself, for example, domain specificity can cause experts to suffer from attentional blindness, missing important information that can result in bias (Dror, 2011). Therefore, as well as understanding what drives face-matching expertise it is also necessary to understand why higher performers, such as face examiners, super recognisers and computer algorithms make errors. By understanding the different sources of error in face matching, the risk of these errors occurring in applied settings, where they can have dangerous and long-lasting consequences, can be mitigated. The following sections provide a critical review of the literature in regard to face-matching expertise, looking specifically at training, face examiners, super recognisers and automated facial recognition algorithms.

## *2.1. Training*

Given the diverse and ubiquitous use of face matching as a means of identification, it is unsurprising that training courses have arisen attempting to improve human expertise in face matching. International guidelines are published online describing in detail relevant training topics as recommended by the face-matching practitioner community. FISWG have published guidance documents on the topic (Facial Identification Scientific Working Group, 2010, 2012b) and the European Network of Forensic Science Institutes (ENFSI) devote an appendix to training in the Facial Image Comparison Best Practice Manual (European Network of Forensic Science Institutes, 2018).

When deployed operationally, automated facial recognition systems seldom work in a 'lights out' capacity (without human intervention), and it is likely to be a long time before humans are no longer required in the face-matching decision process. Spaun (2009) advocated the necessity of human face-matching training, foretelling the increased use of automated facial recognition systems and the requirement for human operators to be the visual arbitrators of the machine's output. The topics recommended by Spaun closely resemble those of FISWG and ENFSI, comprising of:

1. 'General knowledge', including the history of facial comparison, biometric advances and the underlying principles of 'photographic' comparison.
2. 'Image Science', including the properties of digital images and cameras, distortions introduced by imaging systems and for advanced training more detailed descriptions of optical distortions, such as lens barrelling and illumination.
3. 'Image processing' including, at the basic level, brightness and contrast adjustments, rotation and cropping. For advanced training this topic includes sharpening and blurring and separation of colour channels

4. 'Facial specifics' including properties of the face, described as the structure of bones and muscles, facial expression, a working knowledge of visible features within the skin, special attention to the characteristics of the ear, face shape and the statistical prevalence of these shapes within the population at an advanced level of training. 'Facial specifics' also covers alteration face, including aging, trauma and transient changes as well as image manipulation.
5. 'Legal issues' relevant to country and region, including case law relating to facial comparison, admissibility of comparison findings as evidence and how to give testimony in court.

The topics listed by Spaun are detailed, appear relevant to face matching and largely reflect the training advised by best practice groups, but at the time of writing the efficacy of these topics in improving face-matching accuracy had not been demonstrated in the scientific literature. In fact, the recommendation to classify subjects in images based on face shape ('4. Facial Specifics') has since been demonstrated to hinder, rather than aid, face-matching performance (Towler et al., 2014). The intention of this approach is to assign specific faces to a category of shape, with the intention of then discriminating between different faces based on what category or classification is assigned. Towler et al. (2014) evaluated the face shape strategy and found face shapes to be neither diagnostic of identity nor to provide any increase in face matching accuracy. Undergraduate students were provided with written instructions in classifying faces based on the following shape categories:

- Oval
- Diamond
- Oblong
- Round
- Triangle/Heart
- Pear
- Square/Rectangular

Repeatability for classifying the shape of the face was found to be low between observers, with each individual face being classified on average as having three different face shapes by different observers. Of most concern, within-observer agreement was also low, only 56% of repeat classifications of the same face by the same observer were consistent. Instruction in face-shape classification had no impact on accuracy, sensitivity or criterion on the Glasgow Face Matching Test (GFMT) (Towler et al., 2014). Ritz-Timme et al. (2011) took this a step further, evaluating the repeatability of classifications for a list of 43 facial features using a published facial feature atlas. The study found inconsistencies in feature classifications between different observers. Even trained and experienced individuals had a mean mismatch percentage of 39% when using the atlas (mismatch percentages ranged from 14%-70%), meaning that two trained operators disagreed on a feature classification for the same face on average 39% of the time. The images used in this study were those used to create the facial feature atlas and thus represent the best possible scenario. It can only be assumed that introducing factors that impact on facial feature shape, such as expression, pose and camera angle, would further decrease consistency (Towler et al., 2014).

In fairness to Spaun, these studies were published after her article and more recent practitioner guidance recommend face shape classification not be used in face matching

(European Network of Forensic Science Institutes, 2018; Facial Identification Scientific Working Group, 2019a). But the lack of empirical validation of face matching training is an issue for most recommended face matching training topics, including facial anatomy and photography training.

### 2.1.1. Facial anatomy and photography training

Practitioner working groups recommend that training in facial anatomy and image capture and processing are required topics for face-matching professionals (Facial Identification Scientific Working Group, 2012b). Lee et al. (2006) evaluated the performance of post-graduate anatomy students considered to be trained in facial anatomy against untrained participants on a low-quality CCTV face-matching task. Anatomy training was found to have no effect on face-matching accuracy. Towler, (2016) assessed the impact of both a 12-week anatomy training course and a 13-week forensic photography training course on face-matching accuracy. Students were tested before and after training. Students on a forensic psychology course, unrelated to face matching, were included in the study as controls. No improvement was found in trainee face-matching accuracy after anatomy and photography training or in relation to the accuracy of controls. A second test using more challenging images also found no overall benefit in accuracy from anatomy or photography training, though results were harder to interpret due to a dissociation in the difficulty of match and non-match pairs on the test. The results suggest that anatomy training may have provided an advantage for match trials compared to controls, but Towler postulated that this could be caused by random noise in the data and further verification of this finding is required.

These studies demonstrated that training and knowledge in facial anatomy and photography alone seems to provide little if any benefit for face matching accuracy. This finding is not that surprising considering the nature of the training under evaluation. The

literature has made it clear that face matching is a difficult task, with nuances of complexity that are very much specific to the face-matching problem. The training assessed in these studies covers topics that relate to two of the major components that underpin the matching of faces in images: facial anatomy (faces) and photography (images). However, they do not address at all the third component: how to compare the images. As well as facial distinctiveness and image quality, the perceptual and cognitive processes that underpin the matching decision are major contributors to accuracy, as shown by the high individual variance in innate ability on face matching tests (Burton et al., 2010; Fysh & Bindemann, 2017b). Given that there are face-matching specific training courses available, a key question is whether these dedicated courses are any better than facial anatomy and photography courses at improving face matching accuracy.

### 2.1.2. Short face-matching training courses

Woodhead et al. (1979) evaluated an instructor-driven person recognition training course that advocated a feature-based approach to face recognition and matching. On a subsequent test, comparing four target faces to a gallery of 240 target faces displayed for a duration of 10 seconds, there was no observable improvement in accuracy after the three-day training course. The validity of this study as an accurate representation of current recommended training practices is questionable, particularly given that the content of the training used in the study predates the publication of international guidelines and best practice in face matching by three decades.

Recently, Towler et al. (2019) evaluated two, one-day online face matching courses. When assessed against FISWG guidelines both courses complied with only 20% of the recommended topics but included components that have been demonstrated elsewhere in the literature to benefit face matching accuracy:

- Encouraging a feature-by-feature comparison strategy also known as the 'morphological' approach (Towler, White, et al., 2017).
- Giving feedback on face matching tasks (White, Kemp, Jenkins, & Burton, 2014).

Participants were tested before and after training on two face matching tests. Towler et al. reported no significant differences in accuracy between pre- and post-training, nor compared to a control group that completed a workplace health and safety course. Based on these findings, short online face matching training courses do not appear to improve accuracy on face-matching tasks.

Online training is not the only delivery method used for face-matching training. Some courses are delivered by a subject-matter expert instructor for longer durations than a single day. For example, courses undertaken by facial reviewers in Australia ranged from one day to two weeks, though one day courses appeared to be by far the most common (Heyer, 2013). In addition to evaluating the two short online courses, Towler et al. (2019) also reviewed two contemporary, professional face-matching training courses delivered by subject-matter experts. The courses were a half-day and three-days in duration. Both instructor-driven courses largely complied with international guidelines and included elements demonstrated in research to improve face-matching accuracy (feedback on comparisons and facial feature comparison). Like the short online training courses, the half-day instructor-driven course provided no improvements in accuracy after training or compared to controls. The findings for the three-day course were more complex as post training improvements were seen for some face-matching stimuli but not others. Where improvements did occur, the effect was small and inconsistent, and for more challenging face-matching stimuli (including incongruent image quality and variability in pose and expression), no improvements in accuracy were observed.

All the course evaluations discussed so far have ranged from one hour to three days in duration, with only the three-day course providing any improvement in face-matching accuracy. As face matching can be considered a challenging perceptual and cognitive task, it perhaps should be expected that such short training would not have much of an effect. Anderson (1982) claimed it takes at least 100 hours of learning and practice to develop a cognitive skill to any level of competency. By this logic, if an individual undertook 6 hours of face-matching learning and practice a day it would take 17 days to acquire any significant proficiency in the task. Seitz & Dinse (2007) cited radiologists as an example where perceptual ability in a visual task is enhanced by extensive exposure or training to stimuli, in this case identifying the presence of a tumour in scans that would be otherwise un-interpretable to the untrained eye. In the UK radiologists receive five years of training, including three years of general radiology training and two years of specialist training in a particular area[3]. The minimum amount of learning and practice needed to improve face-matching ability has not been established and if there is such a cut-off it would likely vary based on the type of task, and the quality and variability of facial stimuli employed. For unfamiliar face learning and recognition (rather than matching) research has shown practice improves the ability to recall faces viewed previously (Hussain et al., 2009a, 2009b). However, any improvements are largely stimuli specific (i.e. the benefits do not transfer to novel face stimuli) (Hussain, McGraw, et al., 2012). Based on these examples it would

[3] https://www.healthcareers.nhs.uk/explore-roles/doctors/roles-doctors/clinical-radiology/training-and-development

appear that to provide any sustained and significant gains in face matching ability that extensive training and exposure to varied face comparison stimuli would be required.

Towler et al. (2019) recommend that given the limited benefits found from short training courses, professional training in face matching should be evidence-based, using techniques and methods that have been empirically demonstrated to improve face-matching accuracy. The following sections review some of the strategies from the literature that have been demonstrated to improve face-matching accuracy in controlled laboratory experiments.

### 2.1.3. Morphological feature comparison

Current best practice from practitioner working groups advocates the morphological comparison approach as the preferred method for unfamiliar face matching (European Network of Forensic Science Institutes, 2018). FISWG describe the morphological approach as; *"the method of facial comparison in which the features and components of the face are compared. Conclusions in relation to similarity or difference are based on subjective assessment, evaluation, and interpretation of observations"* (Facial Identification Scientific Working Group, 2019a). Morphological comparison should not be confused with the classification of features discussed previously. The morphological approach is based on the comparison of the shape and form of features rather than comparison of what pre-defined shape category a feature is believed to fall into. The literature suggests that cognitive processes used to compare unfamiliar faces share commonality with those used in visual pattern matching rather than the more specialised holistic processing used for familiar faces (Megreya & Burton, 2006). On this basis the use of a feature-based comparison process sounds logical. Towler et al. (2019) found that of eleven face matching training courses reviewed, all taught a morphological, feature-by-feature comparison approach.

When two groups of undergraduate students completed a face-matching task by comparing individual features or holistically (i.e. comparing the face globally and not using individual features), those in the feature group showed no significant improvement in accuracy after receiving this instruction. The holistic group showed a reduced response time but also a reduction in accuracy (Megreya, 2018). Although the students in this study did not demonstrate an improvement, other studies have found some benefit from employing a feature comparison approach. Using the same images Megreya & Bindemann, (2018) asked three groups of undergraduate students to each focus upon one specific facial feature (either the eyebrows, the eyes or the ears) when comparing facial images. They found varied results depending upon what feature the students were asked to focus on. Focusing on eyebrows gave a significant increase in accuracy for match pairs only, no significant change was observed for the eyes group and overall accuracy actually decreased for the group focussing upon the ears. Trained face examiners consider the ears to be one of the most useful features for matching unfamiliar faces (Towler, White, et al., 2017) but Megreya & Bindemann's (2018) study found quite the opposite. Possible explanations for this might be that the images used in the study had been cropped to remove the background and this cropping may have altered the shapes of the ears. The images were also of the front of the face limiting the visibility of the ears. It may be that being able to compare ears effectively requires training and experience, which was not addressed in this study. Eyebrows on the other hand are easily visible in frontal images and would not be affected by cropping. It is likely that what features are useful for matching depends upon the nature of the images being compared, idiosyncrasies in a person’s appearance and possibly the ethnicity of both the observers and the subjects depicted in the images. In Megreya & Bindemann's (2018) study the benefits of using the eyebrows did not transfer to other ethnicity faces but were repeatable for faces of the same ethnicity taken on different days (the images used in the first part of the study were faces of the same ethnicity to the observers and taken on the

same day). It is not clear if the absence of a transfer effect was caused by the other-race effect or the eyebrow being less discriminatory for the other ethnicity faces.

The findings of Megreya & Bindemann (2018) offer evidence that a feature-based comparison strategy can improve accuracy in face matching, but importantly only under certain conditions and only for match pairs. In applied settings where imaging conditions are uncontrolled, focussing only on one feature would be a limiting factor as that feature may be poorly represented in certain images, could be altered or may not even be visible. Operationally, facial reviewers and examiners often have a list of facial features on which to base their decision. FISWG have developed a detailed list of facial features that breaks down the face into 19 different components (see Table 1). These components are then sub-divided into feature characteristic descriptors (Facial Identification Scientific Working Group, 2018). Table 2 shows an example of the characteristic descriptors of the nose.

*Table 1 – List of FISWG feature components*

| ID | Facial Components |
|---|---|
| 1 | Skin |
| 2 | Face/Head Outline |
| 3 | Face/Head Composition |
| 4 | Hairline/Baldness Pattern |
| 5 | Forehead |
| 6 | Eyebrows |
| 7 | Eyes |
| 8 | Cheeks |
| 9 | Nose |
| 10 | Ears |
| 11 | Mouth |
| 12 | Chin/Jawline |
| 13 | Neck |
| 14 | Facial Hair |
| 15 | Lines |
| 16 | Scars |
| 17 | Facial Marks |
| 18 | Alterations |
| 19 | Other |

*Table 2 – FISWG characteristic descriptors of the nose*

| 9 - Nose | |
|---|---|
| Component Characteristics | Characteristic Descriptors |
| 9.1 Nasal Outline (Profile and Front view) | Overall Shape<br>Length and/or width relative to rest of face<br>Prominence<br>Symmetry |
| 9.2 Nasal Root (Bridge) | *Front View: w*idth, length, shape, depth<br>*Profile View:* length, depth, angle |
| 9.3 Nasal Body | *Front View: w*idth, length, shape, angle<br>*Profile View:* length, angle, contour |
| 9.4 Nasal Tip | Shape (in front and profile view)<br>Angle (e.g. up , down)<br>Symmetry |
| 9.5 Nasal Base | Width<br>Height<br>Deviation to the right or left |
| 9.6 Nasal Base: Alae (Wings of Nose) | Thickness<br>Symmetry<br>Shape |
| 9.7 Nasal Base: Nostrils (Nasal Openings) | Shape and size of opening<br>Symmetry<br>Hair |
| 9.8 Nasal Base: Columella (Soft Tissue between Nostrils) | Width and length<br>Relative position<br>Symmetry |

Towler, White, et al. (2017) used a variant of the FISWG feature component list in their study evaluating the impact of a feature-by-feature approach on a quick decision face-matching task. Untrained students that used the facial feature list performed significantly better than students who were not given a list, but again this increase in accuracy was only for match pairs. Students that had used the list then completed the GFMT without using a facial feature list. This group showed no improvement in accuracy compared to the group that had not previously used a feature list, thus there was no transferable training benefit from having used a facial feature list. Interestingly, a detailed analysis of each of the features used on the list showed that the ears were the feature most diagnostic of identity and thus the most useful feature for comparison. This is in contradiction of the findings of

Megreya & Bindemann, (2018), who found that using specifically comparing the ears actually hindered accuracy. This disparity in findings supports the notion that the features useful in a comparison are dependent upon the conditions of the images being compared (e.g. pose, expression, lighting, resolution) and likely vary by individual face.

Taken together these three studies provide an interesting, if sometimes contradictory narrative on the morphological featural comparison of faces:

- Simply asking individuals to use a feature-by-feature approach has no impact on face matching accuracy (Megreya, 2018).
- Asking individuals to focus on certains feature has improved accuracy for match pairs depending upon the feature used, whereas focussing on other features lowered accuracy (Megreya & Bindemann, 2018).
- Providing individuals with a facial feature checklist has improved accuracy for match pairs, which includes features that were detrimental to accuracy when used individually (Towler, White, et al., 2017).

Even if a featural-based comparison strategy improves matching accuracy for unfamiliar faces, making effective use of it is not a trivial matter. The mechanisms by which we match unfamiliar faces may provide a possible explanation for the findings of these three experiments. Although more challenging than familiar recognition, unfamiliar face matching can be performed very quickly in an almost autonomous manner but often not to the same level of accuracy. Kahneman describes these automatic and effortless decisions as 'system 1', where our cognitive processes *'operate quickly, with little or no effortful control'* (Kahneman, 2011, p. 20). Decisions that require more cognitive resource cannot be performed in this autonomous manner and are described as arising from 'system 2'; where

the brain *'allocates attention to the effortful mental activities that demand it, including complex computation'* (Kahneman, 2011, p. 21).

Face matching decisions can be made very quickly with little effort, akin to a 'system 1' process. Fysh & Bindemann (2017c) found that only time pressures under 4 seconds impaired accuracy for comparison of non-match pairs and time constraints had little effect for match pairs. An earlier study found that a time-based impairment to accuracy was most noticeable when decisions were made in less than 2 seconds, whereas lowering decision times from 10 to two seconds had only a slight effect on accuracy (Bindemann et al., 2016). If we are presented with two facial images we will very quickly reach a decision as to whether we think they are the same person or not, with little conscious effort. If this quick face matching process is likened to a 'system 1' decision, then it is unlikely that the process can be wilfully turned off (Kahneman, 2011), which goes someway to explaining Megreya's (2018) findings. Simply telling participants to compare faces feature-by-feature was insufficient to override their quick, almost-autonomous face-matching process. Megreya & Bindemann's (2018) instruction to focus upon a specific feature may have consciously changed the participants face-matching behaviour but conversely may also have caused them to discard a large amount of useful information from other features of the face. This study provides evidence that top-down attention to a specific feature changed face matching behaviour, unlike Megreya's (2018) findings where simple exposure to face matching stimuli without attention to a specific feature had no effect. Top-down attention has been shown to have a significant role in perceptual learning for other sensory tasks (for an overview see Tsushima & Watanabe, 2009).

Faces contain a huge amount of detail. By focussing solely on the eyebrows, other facial features that may provide identity relevant information may be discarded. Towler, White, et al. (2017) addressed this by providing a facial feature checklist aid to record findings,

reducing the demand on working memory and improving accuracy, but again only for match pairs. When the checklist was removed the accuracy gains were no longer observed and participants may have reverted back to a quick intuitive face-matching process. Results also demonstrated an increase in sensitivity when using the facial feature checklist, indicating that the checklist provided a genuine improvement in accuracy for match pairs and was not caused by altering response bias. A small group of face examiners also took part in the study (n=7). Examiner performance using the checklist was not only significantly better than the untrained group, their feature similarity decisions were more diagnostic of identity than the untrained group, notably the examiners made much greater use of scars and blemishes and ears compared to untrained participants. This indicates that examiners had greater knowledge of what features were the most useful for comparison from the checklist and adopted a qualitatively different approach to the novice group.

In summary, employing a morphological, feature-by-feature comparison approach can improve accuracy for match pairs, but which features are most useful depends on the images being compared. To more fully utilise the large amount of feature information available in a face, some kind of decision aid is necessary, in this case a facial feature checklist. Findings also suggest that the training and experience of face examiners improves their accuracy when using a morphological, feature-by-feature approach.

### 2.1.4. Feedback training

Perceptual learning is broadly defined as improvement on a perceptual task that is induced or facilitated by practice or experience (Tsushima & Watanabe, 2009). Feedback on tasks is a key component to inducing perceptual learning, with improvement on some perceptual tasks being induced solely by feedback (Choi & Watanabe, 2012). Alenezi & Bindemann, (2013) administered feedback on responses over a prolonged face-matching task using

images from the Glasgow University Face Database. 200 face image pairs (100 match pairs and 100 mismatch pairs) were divided into 40 image pair blocks. Participants in the feedback condition received an initial block without feedback, then three feedback blocks followed by a final three blocks without feedback, including previously seen and novel face stimuli. A second group completed the same seven blocks but without any feedback. Feedback failed to provide any improvements in accuracy for either old or new face image pairs but did prevent a decline over time in accuracy on mismatch pairs, observed in the non-feedback group. These findings contradict those of White, Kemp, Jenkins, & Burton, (2014) who also investigated feedback training for face matching, giving untrained participants immediate feedback on a training test of 168 facial image pairs from the GFMT (the same images used by Alenezi and Bindemann). Participants then completed a more challenging transfer test of novel stimuli without feedback, consisting of 80 match and 80 mismatch facial image pairs. A control group completed the same facial image pairs but without any prior feedback training. Participants were split into low aptitude and high aptitude groups based on a pre-screening face matching test. Feedback was observed to provide a significant improvement for the low aptitude group on both the training test and the transfer test. The authors suggest that after feedback participants were more aware of the features that reliably predict identity and this transferred onto new stimuli, providing perceptual learning through both feedback and attention to features, however further testing is required to confirm this hypothesis. It is important to note that the perceptual learning benefit was only observed for lower aptitude participants whereas high performers were not improved by feedback.

How the feedback was presented may account for the different findings from these two studies. White et al. presented corrective feedback whilst the faces remained on the screen, allowing participants to review their decision and potentially learn what features were

reliable or unreliable. Alenezi and Bindemann presented feedback without the faces onscreen preventing any opportunity of reviewing the images. Although the two studies differ in their findings both show a benefit from the administering of feedback during training:

1. Preventing a decline in accuracy for mismatch facial image pairs (the inability to tell faces apart).
2. Improving accuracy for low aptitude performers on novel facial image pairs.

Of the two methods for delivering feedback, providing feedback alongside the facial image pair appears to be the most beneficial. White et al. suggest that increasing the diversity of stimuli in training sets may increase the transfer benefit for novel stimuli, as has been observed in other visual perception tasks (Hussain, Bennett, et al., 2012).

Feedback has also been shown to benefit novices in fingerprint comparison, but in a different manner to that seen for the two face-matching studies. Searston & Tangen, (2017a) tested three strategies in an attempt to induce perceptual learning for novices conducting fingerprint comparisons:

1. Providing immediate feedback on comparison decisions.
2. Encouraging participants to generate label descriptors to characterise observed similarities and differences between a pair as a method of 'elaborative interrogation' (generating learning through the explanation of an answer). This is similar to the facial feature checklists used by Towler, White, et al. (2017).
3. Introducing contrast into the comparison decision by providing both matching and non-matching exemplars to the target, theorised to help trainees learn to contrast within- and between-subject variance.

A second group completed the fingerprint comparisons following a baseline training protocol. Results demonstrated that each of the three conditions produced perceptual expertise in a fingerprint matching, but only for mismatch pairs. All three training conditions had no significant impact upon accuracy for match pairs but did increase sensitivity for features that discriminate highly similar mismatches. For the feedback and labels condition this was a significant finding. The findings of the labels group are in direct contradiction to those of Towler, White, et al. (2017) when using facial feature checklists. Use of facial feature checklists gave greater accuracy on match pairs rather than mismatch pairs, whereas labelling features for fingerprint comparisons did the opposite.

Gentry & Bindemann (2019) provided untrained participants with example matching and non-matching face pairs when comparing images from the GFMT, similar to the contrast condition used by Searson & Tangen (2017a). Face-matching examples significantly improved the accuracy of low performing participants on match and non-match trials and the effect persisted for new images from the GFMT shown without examples. However, the effect was not observed for new face-matching stimuli from another test. Gentry & Bindemann believe that the examples may have helped the low performing participants to form a more consistent decision criteria for the GFMT images but this decision criteria did not generalise to the new, more variable images from the Kent Face Matching Test (KFMT).

There are a number of fundamental differences between Searston & Tangen's (2017a) study and the studies by Towler, White, et al. (2017) and Gentry and Bindemann (2019). Firstly, the facial feature checklist used by Towler, White, et al. (2017) imposed labels whereas Searston & Tangen (2017a) allowed participants to generate their own labels. Secondly, labelling fingerprints improved accuracy on non-matching trails whereas the facial feature checklist improved accuracy on match trials. Thirdly, Gentry & Bindemann provided labelled matching and non-matching examples, whereas in Searston & Tangen's

study participants had to decide which of the contrasting fingerprints was a match and which was not a match. Of course, the most fundamental difference between the studies is that one used fingerprints as stimuli and the other unfamiliar face images. This warrants investigation of the feedback, labels and contrast protocols used by Searston & Tangen when applied to facial images. Of particular interest is whether the combination of the feedback, labels and contrast could provide an additive effect of perceptual learning for face matching.

### 2.1.5. Within face variation

Part of the challenge of face matching in applied contexts is derived from the limited visual stimuli available, as observers are restricted to the images being used to make the match decision and cannot rely on the multiple and diverse representations that would be available if the image was of a familiar face. As well as struggling to tell faces apart, we can also find it difficult to tell faces together (Young & Burton, 2018). If two people appear sufficiently similar they may be mistaken as the same person under similar imaging conditions. Varied imaging conditions can also affect images of the same person, making them appear sufficiently dissimilar so that they may be mistaken for images of different people. When tasked to sort multiple images into constituent identities participants consistently overestimated the number of identities, believing there to be more individual identities present than there actually are (Andrews et al., 2015; Jenkins et al., 2011; Sauerland et al., 2016). These identity-sorting errors are understood to be caused by participants misinterpreting the *within-person variability* of unfamiliar faces as *between-person differences*. Informing the participants of the correct number of constituent identities greatly increases accuracy on an identity sorting tasks (Andrews et al., 2015).

Approaching the problem of face variability from a more applied perspective, Bindemann & Sandford, (2011) demonstrated that by providing additional images of the same person to compare to multiple targets on an unfamiliar face matching task, accuracy could be significantly improved. With only a single image to compare, mean accuracy on the task was a meagre 57%, but when participants were given three images to compare to the targets, accuracy rose to 85%. White, Burton, et al. (2014) also found that increasing the number of images to compare to a target face increased accuracy, but only for match trials. As providing additional images did not impair performance on mismatched trials, sourcing additional images of an individual when comparing facial images may be a worthwhile means of increasing accuracy on match trials.

### 2.1.6. Mentoring

The research on training thus far has provided limited evidence that professional training courses improve face-matching accuracy, but highly trained and experienced face examiners do, as a group, perform better than untrained individuals. One possible contributing factor to the enhanced accuracy of examiners that has not yet been explored is mentoring. Training for forensic experts is believed to be much more in depth than a one- or two-day training course. Current best practice recommends forensic trainees complete detailed and tailored training programmes, often with an assigned mentor to guide them through the process (European Network of Forensic Science Institutes, 2015, 2018). Tacit knowledge in the context of perceptual tasks can be very difficult to describe to novices in manuals or formalised training procedures (Engstrom, 2003) and mentoring within an organisation can be an effective mechanism for the transfer of valuable tacit knowledge to trainees (Swap et al., 2001). It may be that mentoring in face matching training programmes provides an additional, effective mechanism for knowledge transfer that cannot be achieved through short formalised training courses. Dowsett & Burton, (2015) investigated the

potential of untrained participants working on face matching tasks in pairs as a route to improving accuracy. They found that, particularly for low performers, working in pairs did improve accuracy on subsequent face-matching tasks completed individually. Although far removed from applied working practices this study does demonstrate that working with others on a face-matching task facilitates knowledge transfer and produces subsequent improvements in accuracy. Further research is required to establish how knowledge transfer through mentoring is, or can be, applied to face matching to drive improvements in accuracy.

### 2.1.7. Training overview

Face matching in applied settings is a challenging task. Fysh & Bindemann (2017a) attribute this challenge to both a *data problem* and a *resource problem*. The data problem refers to the limitations of the images being compared, including image quality, number of images available and within-face variability. The resource problem refers to the observer making the match and their working environment, including individual face-matching ability, response biases, other inherent biases (e.g. own-race bias) and time pressure. Ideally, effective face matching training should address all of these problems. Of the research that has been conducted on face-matching training effectiveness, the results are at best limited. For training strategies that do appear to show a benefit, often these are marginal, apparent only for individuals with initially low levels of face matching ability or provide improvements for matching facial image pairs but not non-matching facial image pairs. Studies also present conflicting results on the benefits of different training strategies depending upon experimental design, including morphological feature comparison (Megreya & Bindemann, 2018; Towler, White, et al., 2017) and feedback (Alenezi & Bindemann, 2013; White, Kemp, Jenkins, & Burton, 2014). Table 3 summarises the research findings to date for face matching training strategies.

*Table 3* – Accuracy improvements from face matching training

| | Accuracy Improvement | | |
|---|---|---|---|
| Training | Overall | Match Pairs | Non-match Pairs |
| Facial anatomy | ×* | ×* | × |
| Photography | × | × | × |
| Short courses (1hr - 1.5 days) | ×** | ×** | ×** |
| Short courses (3 days) | ✓** | ?** | ?** |
| Longer courses (>3 days) | ? | ? | ? |
| Morphological comparison | ✓ | ✓ | × |
| Feedback | ✓ | ✓ | ✓ |
| Examples | ✓ | ✓ | ✓ |
| Within-face variability | ✓ | ✓ | × |
| Working in pairs | ✓ | ?*** | ?*** |

*Match accuracy improvements were observed in one study but may be caused by noise in the data and require further examination (Towler, 2016).
**Only limited and inconsistent improvements were observed from a three-day instructor led course. Shorter course gave no improvements (Towler et al., 2019)
***Working in pairs provided an accuracy improvement but it is not known if this improvement was for both match and non-match pairs (Dowsett & Burton, 2015).

Validating an aspect of training in isolation is important. From a theoretical view it demonstrates whether a particular training strategy affects the cognitive processes that underlie face matching and for the applied users it is useful for confirming what aspects of a training course are effective. However, this piecemeal approach is quite far removed from applied practice and may overlook the complexities of training as it is delivered professionally or if there are interdependencies that exist between different training strategies. Studies of professional face matching training courses in their entirety are limited, but the results are not promising as an effective route to expertise. Towler et al. (2019) demonstrated that contemporary professional face matching training courses provide little, if any, benefit to face matching expertise when training courses are three days or less in duration. Based on the limited effects and contradictory findings from the literature it would appear that the training practices evaluated so far do not provide an effective route to face matching expertise.

There is a possible exception to this trend, which is the enhanced performance of trained forensic face examiners. According to working documents from practitioners examiners undergo long durations of training and mentoring (European Network of Forensic Science Institutes, 2018; FISWG, 2019), which may contribute to their enhanced face matching ability. To date there has been no longitudinal study of face examiner training to confirm if this is the case. Also, the training practices of different organisations that employ face examiners are not well understood outside of the face examiner community, barring high-level guidelines regarding training topics from practitioner working groups. Longitudinal training studies and an in-depth investigation of examiner training practices are warranted to determine whether training does indeed drive examiner expertise, as thus far research into short professional training courses has not been fruitful.

## *2.2. Forensic face examiners*

Forensic face examiners, or face examiners for brevity, are trained face-matching professional that commonly work within small specialist teams, providing expert testimony for court or as an escalation point for challenging face-matching tasks. Face examiners have presented evidence in UK courts over 27 years, with the first stated court case in 1993 where the practice is colloquially referred to as 'facial mapping' (*R v Stockwell*, 1993). Forensic face matching evidence is commonplace in UK criminal trials and in 2003 it was estimated that approximately 600 face matching examinations were presented in court a year in England and Wales (Bromby & Plews, 2003).

Surprisingly, despite the growing use of face-matching evidence from face examiners there has been little research into face examiner expertise or how examiners carry out face-matching examinations. Where guidelines and best practices do exist, they are vague and not prescriptive about how examinations should be conducted (Steyn et al., 2018). Current

international guidelines recommend that face examiners use a morphological comparison approach for face matching, whereby the face is compared feature-by-feature within a series of stages known as ACE-V (analyse, compare, evaluate and verify) (European Network of Forensic Science Institutes, 2018; Facial Identification Scientific Working Group, 2019a). However, how examiners actually do this in practice is not widely understood nor is the consistency between different examiners in their approach to face matching.

Recent studies have demonstrated that trained forensic face examiners have enhanced face-matching abilities at the group level (for an overview see White et al., 2021). This has been shown for both quick decision face matching (White, Phillips, et al., 2015) and when using face-matching examination procedures (Phillips et al., 2018), referred to respectively as perceptual skill and operational accuracy by Towler et al. (2018). There is also evidence that face examiners demonstrate some hallmarks of face-matching expertise. When matching faces using a checklist of facial features, the similarity ratings of individual features by examiners were more diagnostic of whether faces were matching or non-matching, compared to similarity ratings of novices (Towler, White, et al., 2017). This indicates that face examiners are able to interpret and match faces more accurately at subordinate levels of detail, using specific facial features rather than the face as a whole. The examiners in this study also showed greater understanding of which facial features were more diagnostic of identity, making greater use of distinctive features such as ears and facial marks. Growns & Martire (2020) refer to this marker of expertise as distributional statistical learning, a key cognitive mechanism for enhanced pattern-matching expertise. Distributional statistical learning provides experts with knowledge of features that are rare or uncommon within their domain, which allows them to more accurately differentiate matching and non-matching stimuli.

Other studies have shown that face examiners are more cautious with poor quality or otherwise challenging face-matching stimuli, where they are less likely to make errors with high degrees of confidence compared to novice controls and super recognisers (Norell et al., 2015; Phillips et al., 2018). This implies that examiners have a deeper understanding of the risks of error when making face-matching decisions. Knowing when not to make a decision due to insufficient information is seen as the crux of expertise across many forensic comparison disciplines (Towler et al., 2018). The trait has been observed for trained experts for several fields outside face matching, including speaker comparison by forensic phoneticians (Bartle & Dellwo, 2015), expert handwriting analysis (Sita et al., 2002) and species identification by wildlife experts (Austen et al., 2016).

Given the apparent limitations of face-matching training, it is somewhat surprising that face examiners outperform novice groups in face matching. Towler et al. (2021) use the term 'training paradox' to refer to the limited efficacy of face-matching training strategies and the apparently contradictory enhanced performance of face examiners. Towler et al. (2021) theorise that the morphological comparison approach used by examiners is a process distinct from the natural cognitive mechanisms humans use to identify faces (see Section 2.1.3 for a more detailed discussion). They suggest that to develop expertise in this approach requires significant cognitive effort to both develop the skill and override the automatic, face processes that occur naturally. Based on the available evidence of face examiner expertise this appears feasible. However, to date there are only seven published studies evaluating the accuracy of face examiners (White et al., 2021) and there has been very little advancement in the theoretical understanding of examiner expertise.

Prior to 2015, much of the research literature published on forensic face examiners was in fact highly critical of examiner working practices. Techniques used by face examiners have been largely borrowed from other disciplines (Bromby, 2006) and subsequent research has

demonstrated that certain techniques have little, if any, scientific validation (Mallett & Evison, 2013; Moreton, 2021). Some are no longer recommended for use in current best practice guidelines (European Network of Forensic Science Institutes, 2018). These techniques include photo anthropometry (comparing faces using measurements and ratios) (Kleinberg et al., 2007; Kleinberg & Vanezis, 2007; Moreton & Morley, 2011), superimposition (overlaying faces to highlight similarities and differences) (McNeill et al., 2015; Strathie et al., 2012; Strathie & McNeill, 2016) and feature classification (assigning facial features to categories based on shape) (Towler et al., 2014). Whilst there is some evidence that the morphological feature comparison approach does improve face matching accuracy, the technique has only been tested on a limited range of facial imagery with a small number of examiners (n = 7) (Towler, White, et al., 2017). Wider testing of the technique with more examiners and different types of imagery is required to establish the reliability of the technique.

The absence of known error rates for face examiner techniques has been highlighted as a major concern (Edmond et al., 2009), and is an issue that is rife across many forensic pattern-matching disciplines (National Academy of Science, 2009). In 2016, the US President's Committee of Advisors on Science and Technology (PCAST) raised the pressing need to scientifically validate forensic feature-comparison methods through black box testing, to establish known error rates and prevent the risk of unreliable examinations entering the court room as expert evidence. Black box studies test forensic examiners with a series of ground truth matching tasks that emulate those encountered operationally. The results of the test are then used to establish how often examiners make incorrect decisions on the task. The intention of widescale black box testing is to measure the operational accuracy of examiners and provide greater transparency regarding error rates on applied tasks. Although black box studies have started to measure both the perceptual skill and

operational accuracy of face examiners (Phillips et al., 2018; White, Phillips, et al., 2015), there is a need for greater transparency regarding individual differences in examiner performance. To do so requires a move towards white box testing, which will improve the theoretical understanding of how examiners approach face matching tasks.

### 2.2.1. Perceptual skill

The first notable study to measure the perceptual skill of face examiners was published by White, Philips, et al. in 2015[4]. 27 face examiners took part in a series of face-matching experiments. At the group level examiners outperformed novice controls on the GFMT and the more challenging Person Identification Challenge Test (PICT). In both tests participants viewed a series of facial-image pairs and were asked to decide if the images in a pair were a match or not a match. Examiners were not allowed to use their standard examination tools and procedures with viewing times restricted to 30 seconds.

In a third experiment facial-image pairs from the Expertise in Facial Comparison Test (EFCT) were shown in two timed conditions (2 seconds and 30 seconds) with half of the images presented upright and half inverted. Again, examiners outperformed novices at the group level in all conditions. The examiner advantage was most pronounced when facial images were shown for 30 seconds rather than 2 seconds, supporting the notion that examiners make more considered face matching judgements compared to novices. Examiners were also less impaired by image inversion, with the authors suggesting that

[4] Wilkinson & Evans published an earlier study in 2009 concluding that 'facial imagery experts' were better than the general public at matching faces, however the two authors were the only facial imagery experts to take part in the study, limiting the applicability of the results to the wider face examiner community.

examiners expertise is feature-based and qualitatively different from the holistic processing used for familiar faces.

The study by White, Phillips, et al. (2015) was one of the first and most comprehensive studies of face examiner perceptual expertise. However, there are a number of limitations to the findings. The authors state that given the experimental constraints of the study the findings are an *estimate* of face examiner expertise to act as a benchmark in further testing. Across all three experiments the examiner group made errors 7% of the time, showing that examiner perceptual skills are far from perfect. Perhaps most importantly, all data were reported as group means with a margin of one standard error from the mean. The performance of individual examiners is not shown, and it is possible that the mean results are masking a wide range in individual examiner performance, as is commonly found for novices in face matching tasks (Bindemann et al., 2012). Although examiners show evidence of enhanced perceptual skill in face matching, to truly understand the extent of this expertise future studies should report individual performance on tasks. Doing so will allow researchers to understand the range in perceptual skill within the examiner community and confirm whether all examiners show similar levels of skill or if group performance is being driven by a smaller number of exceptional individuals.

### 2.2.2. Operational accuracy

Phillips et al. (2018) published the largest study of face examiner operational accuracy to date, in direct response to calls for independent and objective research into forensic pattern-matching disciplines by the 2009 NAS report. 57 examiners recruited across five continents took part in the study. Where White, Phillips et al. (2015) focussed on perceptual skill in timed conditions, this study allowed examiners access to their tools and procedures and gave them up to three months to complete the study. The only constraint was that

responses must come from an individual examiner rather than a group or aggregate response. The task consisted of 20 challenging face-matching pairs and performance was reported using the area under the receiver operating characteristic curve (AUC). AUC measures the extent to which a similarity/dissimilarity judgement predicts the class of a set of stimuli (Towler, White, et al., 2017). In this case participants rated whether facial images were a match or non-match using a seven-point Likert scale, ranging from -3 for non-matching pairs to +3 for matching pairs.

Face examiner performance on the task was compared to two other face specialist groups (facial reviewers (n = 30) and super recognisers (n = 12)), fingerprint examiners (n = 53), novice undergraduate students (n = 31) and four state of art automated facial recognition algorithms. The median AUC score of examiners outperformed all other human participant groups and three of the four algorithms. For human participants the median difference in AUC between face examiners was significant when compared to fingerprint examiners and undergraduates, with 53% of face examiners performing above the 95$^{th}$ percentile of the student AUC distribution. Unlike White, Phillips, et al. (2015) this study does report individual AUC scores for each participant. For all groups individual performance varied widely, including face examiners. Seven out of the 57 face examiners achieved a perfect AUC of 1 but all others made errors and most worryingly, six examiners performed more poorly than the average undergraduate student. These findings reveal that, concerningly, the enhanced performance of examiners at the group level is driven by a proportion of exceptional individuals rather than a more general enhanced performance of all the examiners, suggesting that examiners are not a homogenous group.

Phillips et al. (2018) concluded that face examiners are more accurate than non-face specialist groups and performance can be further increased by crowdsourcing or fusing responses. However, the study seems to overlook the potential implications of the drastic

variation observed in individual face examiner performance. For end-users of face examiner decisions, such as a police investigator or the courts, the fact that some examiners seemingly perform so poorly at face matching is a major concern. Black box studies are helpful in demonstrating the variation of performance in examiner groups, but by their nature do not provide insights into why variation exists and do little to further the theoretical understanding of examiner expertise. The enhanced performance of face examiners at the group level is a promising finding but more research is needed to understand how examiners make their face-matching decisions. To do so requires the face examiner community to participate in white box testing.

### 2.2.3. Face examiner expertise

Black box tests are informative in terms of how well face examiners perform at face matching tasks but provide little if any information on how they conduct the task. Understanding how forensic experts carry out visual comparison tasks is generally lacking across forensic science. Research into the psychological processes that expert examiners use to make their decisions is beneficial for unpacking why exceptional examiners perform well, leading to the development of more effective training and development programmes for the examiner community (Growns & Martire, 2020b). Much of the evidence for face examiner expertise rests on the use of a feature-based, morphological approach, with studies demonstrating that face examiners perform better at face matching with more time (White, Phillips, et al., 2015) and are more proficient than novices when using a facial feature checklist (Towler, White, et al., 2017). It is interesting that this finding implies face examiner decisions are deliberate and require conscious effort whereas in many other domains automaticity in decision making is a hallmark of expertise (Edmond et al., 2017). In reality, the division between automatic ‘system 1’ and deliberate ‘system 2’ decision making may not be so clear cut and face examiners may use a hybrid approach, for example

a cut down analytical approach when under time pressure (Growns & Martire, 2020b). Or conversely, face examiners decision making process may be more automatic than they are consciously aware of and the detailed comparison of individual facial feature are little more than retrospective attempts at rationalising a cognitive process that is largely inaccessible. The use of retrospective rationalisations by experts in other domains is well documented in the literature (see Edmond et al., 2017). Preliminary findings that examiner facial feature similarity ratings are more diagnostic of whether a face pair are a match (Towler, White, et al., 2017) indicate the morphological approach is used in examiner decision making. However, more research is needed to uncover the cognitive processes examiners use to make face-matching decisions in security critical and high-risk applied settings.

Studies of forensic fingerprint examiner expertise provide a useful analogy to forensic face matching and highlight potential avenues for future research. Like face examiners, fingerprint examiners undergo training and mentoring to develop perceptual expertise in a visual matching task. Fingerprint examiners must also match stimuli across a wide range of conditions, from high quality ten-prints to distorted latent crime scene prints. Similarly to face examiners, fingerprint examiners outperform novices when matching under time pressure and in untimed conditions, where they apply a more featural, analytical approach (Growns & Martire, 2020b). A large-scale black box study of 169 fingerprint examiners matching 100 fingerprint pairs of varying quality was conducted by Ulery et al. (2011). Only five examiners wrongly matched a non-matching fingerprint pair, giving a false positive rate of 0.1%. 85 examiners made at least one false negative error giving a higher rate of 7.5% for false negatives. All false positives and the majority of false negatives were detected if reviewed by a second examiner. Examiners were found to disagree most frequently on whether a fingerprint contained sufficient information for examination. A follow up black box study found that when 72 examiners were asked to examine the same stimuli seven months

later, 89% of matching decisions and 91% of non-matching decisions were consistent. Where examiners did change their decision, this was mostly to inconclusive (Ulery et al., 2012). A series of white box studies have since been carried out to understand why fingerprint examiners can reach different decisions and make errors on certain fingerprints. These studies have found that whilst most fingerprint examiners are highly accurate, reproducibility was lowest for deciding whether a print is considered sufficient for examination (Ulery et al., 2013) and examiners differ widely on how many minutiae (features) they consider sufficient for deciding on a match (Ulery et al., 2014). Where fingerprint examiners do differ on their marking up of minutiae on fingerprints this is most pronounced for areas of low visual clarity (Ulery et al., 2016). Investigation into the higher prevalence of false negative errors found erroneous decisions to be generally caused by either challenging stimuli or inappropriate decision criteria for non-matching pairs. Challenging stimuli resulted in misinterpretation of distorted features and fingerprints in different orientations. Examiners were also found to base non-match decisions on a single non-matching feature despite high numbers of matching features being observed, which the authors refer to as the inappropriate use of the “one discrepancy” rule for non-match decisions. Recent research by Robson et al. (2020), contrasting fingerprint feature selection by examiners with that of novices, found a significant difference in feature selection between the groups and greater consistency in feature selection between experts. However, the difference in feature selection was highly dependent upon the clarity and salience of features on a case-by-case basis. Robson et al. (2020) highlight the need for perceptual training of examiners to highlight useful features across a wide range of different stimuli, to encourage meaningful statistical learning. This is a useful example of where white box testing of examiners has led directly to recommendations for improving training and working practices in the examiner community.

To date no study has provided comparable data or insights into the working practices of face examiners. Even where examiners have been tested, false positive and false negative rates are not reported and no study has addressed intra-examiner repeatability or inter-examiner reproducibility. Face examiners have shown enhanced perceptual skill and operational accuracy at the group level, demonstrating that at least some examiners possess face matching expertise. But there is a pressing need for further research to truly understand the strengths and limitations of facial examination in applied settings.

## *2.3. Super recognisers*

Individual differences in ability have been widely reported in the face-matching literature (see Lander et al., 2018). Bindemann & Burton (2021) theorise that these individual differences arise from the complex cognitive strategies and components used to perceive and compare faces, including how we direct attention to faces, perceive differences or similarities in facial features, evaluate these observations and finally decide whether two faces are the same or not. Given that face-matching ability is highly variable and assuming that this ability is normally distributed in the population, it is not a radical leap to assume that some people will naturally be very good at matching faces.

Following on from their earlier research on developmental prosopagnosia, a condition affecting approximately 2% of the population and marked by exceptionally poor face recognition ability, Russell et al. (2009) investigated whether there could be a contrasting group of people who had exceptionally good face recognition ability. They referred to these exceptional performers as super recognisers (SRs). Four SRs who self-reported being very good at remembering and recognising faces took part in three psychometric tests of face perception and recognition. The Before They Were Famous Test (BTWFT) asks participants to identify faces of well-known celebrities before they became famous and is

aimed at testing long term familiar recognition. The Cambridge Face Memory Test (CFMT) targets short term, unfamiliar face memory by tasking participants with picking from an array of faces the correct match to a face briefly viewed beforehand. The long form of the CFMT includes additional items that are made much more challenging by the addition of artificial, visual noise. All four SRs performed significantly above the control mean and three SRs outperformed all of the control participants. On the CFMT all SRs outperformed controls and three achieved a score two standard deviations (SD) above the control mean. Developmental prosopagnosics are diagnosed using the CFMT when performing two SDs *below* the control mean. Based on this cut off, performance 2 SDs above the mean has been quoted as a selection criterion for SRs in the literature (Bobak, Pampoulov, et al., 2016), but in practice there has been a fair degree of leniency in its application (e.g. Davis et al., 2019). In order to test face perception skills that do not rely on memory, Russell et al. also used the Cambridge Face Perception Test (CFPT). In this task participants must match a face to the most visually similar target in an array of six simultaneously presented faces. The six targets have been manipulated by morphing with six other identities to varying degrees. The four SRs performed significantly better than controls at the group level, but with a smaller effect size than for previous tasks and several controls performed at comparable levels to individual SRs. These results provide the first evidence of heterogeneity in SR performance across different face identification tasks. However, subsequent research has criticised the CFPT as lacking the sensitivity to distinguish high performers and not being representative of a real-world face processing task (Bobak, Pampoulov, et al., 2016).

Since the first use of the term by Russell et al. in 2009, researchers have been increasingly interested in the use of SRs to conduct face matching and recognition tasks in applied settings, as shown in Figure 2 by the increasing number of publications on super

recognisers and similar terms since 2009. Given the difficulty of face matching in general and the limited benefits of short training, the deployment of human operators who are naturally superior at the task provides a seemingly simple solution to solving the issue of poor accuracy in applied face-matching tasks (Bobak, Dowsett, et al., 2016).

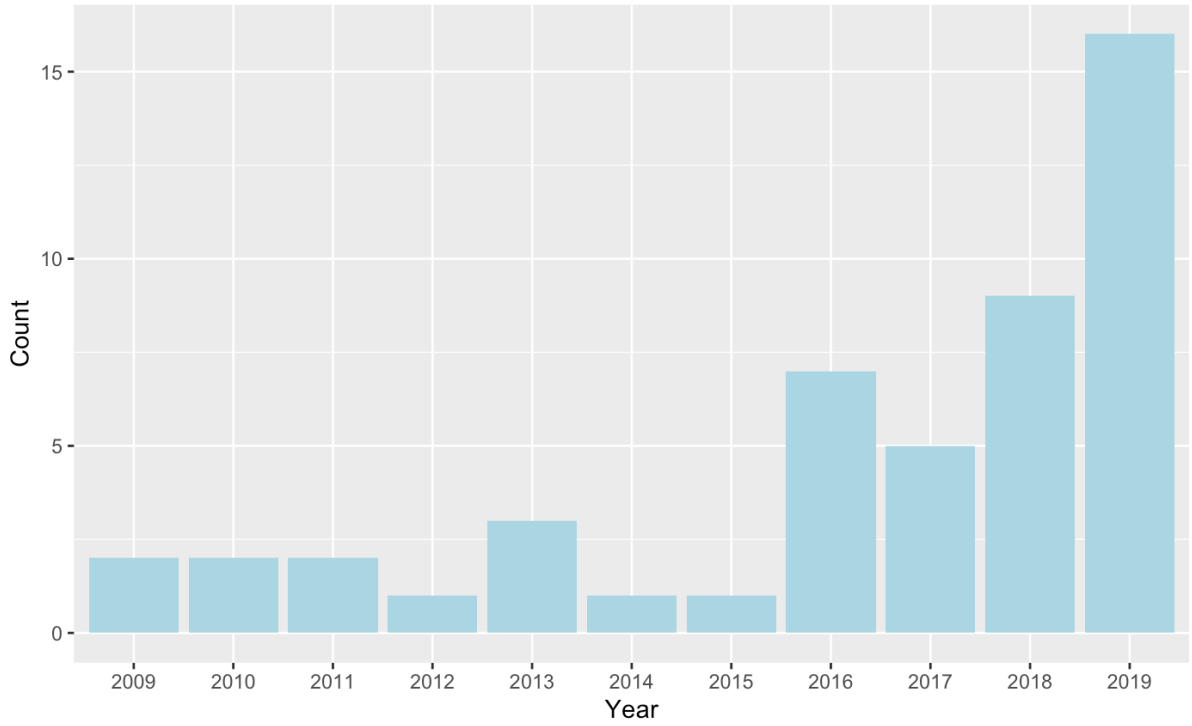


*Figure 2 – Publications by year on super recognisers and related terms*

Media interest in SRs escalated after the much-publicised use of police officers to identify suspects from images and videos as part of Operation Withern, the Metropolitan Police Service's response to the 2011 London riots (Davis, 2019). As much of the imagery used in the investigation came from poor quality CCTV systems, most was unsuitable for matching by automated facial recognition technology at the time. Instead, officers who were believed to be SRs based on either self-reporting or by having made many operational identifications, reviewed the imagery to identify suspects. The high identification rate and subsequent high conviction rate was lauded in the media as a great success in the

operational use of SRs (Davis, 2019). Advocates of SR use have since been unequivocal in the potential benefits of deploying SRs in applied settings, based on operational success stories like Operation Withern (Davis & Robertson, 2020). But despite over ten years of active research, the understanding of SRs and why they are superior has progressed surprisingly little since the seminal paper by Russell et al. in 2009. As a result, applied interest in the deployment of SRs is expanding at a far greater rate than the scientific understanding of their abilities and limitations (Ramon et al., 2019b).

The limited progress in understanding why individuals have superior face processing abilities has led to a *"theoretical vacuum"* in SR research (Noyes et al., 2017). Russell et al. (2009) concluded that their study showed no evidence of SRs being qualitatively different in face processing compared to the normal population but demonstrated that face recognition and perception spanned a far greater range of ability than previously understood. SRs were also shown to vary in individual ability and across different face tasks. Eight years later, based on a comprehensive review of the existing literature, Noyes et al. (2017) similarly concluded that SRs are most likely the top performers from a normal distribution of face processing ability, rather than a distinct group. This is because SRs show consistently high performance at the group level but the performance of individual SRs is variable and does not always exceed control levels. SRs also have a diverse cognitive profile, similar to that of prosopagnosics, with individual SR performance varying on face memory and face matching tasks (Bate et al., 2018), something alluded to by Russell et al.. This indicates that, like prosopagnosics, SRs are an extreme of the face recognition ability spectrum (in this case the upper extreme) and not a distinct group (Bate & Murray, 2017).

### 2.3.1. Group versus individual face matching performance

Several studies have investigated SR performance specifically on unfamiliar face matching tasks and have demonstrated that, while SRs consistently outperform controls at the group level this masks individual differences in SR face matching accuracy. Robertson, et al. (2016) tested the face matching ability of four SRs working for the Metropolitan Police Service using the GFMT short form. The selection process for the SR officers is not given in the study. All four SRs performed better than the control mean but only three exceeded one standard deviation above the mean accuracy of 194 police trainee controls. Due to ceiling effects on the GFMT it is not possible to score two SDs above the control mean, the suggested criteria for SR performance, however one of the SRs did score 100%. On the more challenging Models Face Matching Task (MFMT) the SR group again consistently outperformed the control mean accuracy, this time from 54 undergraduate students (control mean = 73.6% SD = 10.9%, SR mean = 90.3% SD = 1.9%). However, individually the SRs varied in accuracy and none exceeded the two SD cut off of 95.4% correct.

SRs showed significantly superior performance at the group level on both tasks, demonstrating consistently high face matching performance, implying expertise in the task. However, performance varied between SRs and all were surpassed in accuracy by at least one control participant. Clearly the SRs tested had an above average perceptual skill in face matching but were not infallible even on a relatively straightforward face-matching task (the GFMT). Assessing SRs at the group level overlooks any individual differences in performance and conflates the performance of SRs in situations where accuracy is driven by one or two high performers, particularly when small sample sizes are small.

In a similar study, Bobak, Dowsett, et al. (2016) tested seven SRs on the GFMT short form and the MFMT. Bobak et al. used modified *t*-tests to compare SR performance on the GFMT

and MFMT at an individual level. D prime, a measure of sensitivity, and criterion (*c*), a measure of response bias, were also used in a signal detection analysis, providing a more in-depth understanding of SR face matching ability than accuracy alone. Similarly to results from Robertson et al. (2016), SRs at the group level outperformed both controls and motivated controls on the GFMT and MFMT, but not all SRs were significantly better than controls at an individual level on both tests. Individual SRs also varied in their response bias, demonstrating that accuracy on match and non-match pairs was as inconsistent as that of the control participants.

Individual differences and the heterogeneity of SR face-matching accuracy observed in both these and other studies (e.g. Bobak, Pampoulov, et al., 2016; Phillips et al., 2018) questions the validity of evaluating SR performance at the group level, if they are indeed just the top performers within the normal population. Noyes et al. (2017) advocated that SR screening and testing should be reported at the individual differences level, avoiding the conflation of high and low performing SRs in group level analysis. This is an important consideration for the applied community when considering the recruitment of high performers using standardised face-matching tests.

### 2.3.2. Super matchers

Another consideration for SR deployment in applied face-matching roles is the type of selection test used. Many studies of SRs have focussed specifically on face memory and often conflate or ignore face matching ability (Bate et al., 2021). Although face-memory and face-matching ability are correlated to some extent, studies on individual difference in SRs have found performance to vary on different face processing task (Bobak, Bennetts, et al., 2016). For example, Bate et al. (2018) tested 200 British Caucasian adults on three face processing tests designed to replicate different applied tasks, namely a face-memory test,

spotting a face in the crowd test and a face-matching test. The study identified 18 individuals who only achieved superior levels of performance on the face-matching task, dubbed ‘super matchers’ by the authors.

Agencies looking to recruit high performers into applied face matching roles should therefore focus on realistic face-matching tests that are sufficiently discriminating to identify individuals with superior perceptual skills. Despite the fact that face-matching ability has been observed to be inconsistent in typical performers on repeated testing, few studies have investigated the consistency of high performers in face matching over time (Bate et al., 2021). Agencies should use repeat testing to ensure consistently superior performance, one of the hallmarks of expertise.

### 2.3.3. Super recogniser selection

The nature of tasks carried out by operational SRs are not widely understood but are believed to be very diverse (White et al., 2021), likely much more so than the tasks of forensic face examiners. Davis et al. (2019) present perhaps the only specific case example of identification by SRs being used in a UK court. SR police officers were recruited to compare a post-mortem facial image to eight target faces, one of which was that of a missing person believed to be the deceased by the family. According to the article, the coroner confirmed the identity of the deceased based on this exercise and other supporting evidence. However, both the prevalence of this approach and the validity of the procedures used are questionable. Firstly, the UK forensic science regulator has stated that whilst the work carried out by police SRs has investigative value, it is not considered a forensic science and is therefore inadmissible as expert evidence under current UK legislation (Forensic Science Regulator, 2018), making it unlikely that this particular use of SRs is commonplace in British courts. Secondly, because the SRs used in the case were screened

only using the GFMT short form, which, due to ceiling effects, has limited sensitivity, the selection process does not robustly demonstrate that the selected police officers have consistently superior face-matching ability, let alone in the matching of post-mortem to ante-mortem images. In humanitarian emergency response, where post-mortem identification is commonplace, face matching is not considered sufficiently reliable as a means identification (Caplova et al., 2017). Generally, both the selection processes and tasks carried out by SRs in applied settings are poorly understood in the literature.

Despite researchers calling on agencies to capitalise on the advantages that SRs can provide in applied settings (Davis & Robertson, 2020), there are very few studies demonstrating that superior performance on laboratory-based face-matching tests does result in real-world performance gains (Ramon et al., 2019a). In order to provide real-world accuracy benefits, the selection of SRs requires ecologically-valid tests that correlate with the complex and challenging tasks encountered in applied settings (Stacchi et al., 2019). Statistical simulations of SR selection using tests with a correlation to real world tasks of .5 showed only modest gains in accuracy of 12% when selecting SRs with a cut-off of two SDs above the control mean (Ramon et al., 2019a). Diminishing returns in accuracy were seen for using less stringent cut-offs, to reflect that fact that SR selection may need to be done from relatively small numbers of candidates in real world settings. Similarly, Balsdon et al. (2018) found 7% improvements in accuracy on a real-world face matching task when selecting high performers using a one SD above the mean selection cut-off.

These modest gains in accuracy through SR selection alone are unsurprising and can be attributable, at least in part, to the reliability of the selection tests used. As well as face-matching accuracy, other factors will affect test performance, such as the participant's engagement with the task, fatigue, emotional state and also a certain element of chance, which can all be considered as sources of error (Young & Noyes, 2019). These sources of

error will introduce variability into participant performance that, particularly on less sensitive or reliable tests, may mask their true ability to some extent, leading to differences in performance when that participant is tested again. Given the interference of these confounding factors on ability, it is therefore statistically likely that an exceptionally high performer one day will perform closer to average on another day. This phenomenon is known as *regression to the mean* (Kahneman, 2011). If, as appears to be the case, the SR term is simply a label for the top performers within the population, then individual performance will vary on a test-by-test basis, due to regression to the mean and confounding sources of error. Therefore, a person who meets the criteria for SR status one day may not do so on the next and likewise, an unlucky SR may be having a bad day and miss the cut-off for selection due to other factors not relating to their actual ability (Young & Noyes, 2019).

Bate et al. (2019) evaluated the consistency in face-matching performance across three calibrated face matching tests of 30 police officers identified as SRs using a 1.5 SD cut off on the CFMT+. Three of the 30 police SRs significantly outperformed controls on all three face-matching tasks and 24 showed significant performance on only one of the tasks. Interestingly, performance on matching and non-matching trials varied substantially from one SR to the next further demonstrating heterogeneity in SR performance. Bate et al. (2019) thus recommend that repeated screening on calibrated face matching tests is required to identify consistently high performers. Applied face matching tasks are highly varied (Moreton et al., 2019) and the associated risks of misses and false alarms will also vary by task and context (Devue, 2019). Therefore, selection criteria should also factor in variation in performance on matching and non-matching trials within the context of the task that the operator will be required to perform. For example, non-matches at the border may be very uncommon but if missed by an operator will result in a person being wrongfully

admitted into a country. Whereas in policing true matches from a live facial recognition system will be less common than non-matches, but if missed a wanted offender will not be apprehended. It is essential that agencies and organisations understand that if they are looking to select high performers in face matching, the actual operational gains in accuracy from selection alone may in fact be modest and individual performance can vary substantially even between matching and non-matching trials.

### 2.3.4. Are super recognisers experts?

Towler et al. (2021) suggest there are two routes to human face matching expertise, one being the featural-based approach used by forensic face examiners (see Section 2.2) and the other being exceptional ability derived from a person's core face recognition system, which is largely untrainable. The second route of expertise fits the description of super recognisers, or more precisely, super matchers, people with naturally high perceptual skill in face matching. But can such individuals be classed as experts? At the start of this chapter an expert was defined as somebody who shows consistently superior performance in a specific task, acquired by repeated practice and experience. Their performance must also be highly reproducible and show a large, reliable difference to the performance of novices. Based on a review of the small but growing literature on super recognisers, it is apparent that individual performance of different SRs can be highly variable and does not always show a large and reliable difference to controls. In certain studies, there do appear to be highly exceptional performers who meet the definition of face-matching expertise, though the prevalence of such individuals even within cohorts of SRs is low (Bate et al., 2018, 2019). In Bate et al. (2019) only three out of 30 police officers who had *already* been pre-screened as SRs showed consistently superior face-matching expertise across three challenging face matching tasks. This is an important point for applied uses of SRs, where the available pool of recruits for face matching roles may be small. The number of police

officers in England and Wales in 2019 was at its lowest level since the early 1980s[5], meaning resources are stretched and the operational deployment of personnel must be carefully considered. Even for a single selection test, applying a cut-off of two SDs above the mean would result in just two or three candidates meeting the criteria from a pool of 100 (Ramon et al., 2019a), and this number would likely decrease further if those individuals were subject to repeated testing (Bate et al., 2019). In reality, there may not be a sufficient pool of candidates to allow for the selection of face-matching experts. Applying less stringent cut offs will allow for selection of a greater number of high performers but the gains in operational accuracy may be diminished (Balsdon et al., 2018; Ramon et al., 2019a) and it is unlikely the selected candidates could all be considered face-matching experts.

Due to a growing body of evidence that SRs are simply the high performers within a normal population, the number of those that meet the definition of expert is likely to be very small. This has implications for applied agencies and organisations hoping to recruit face-matching experts into operational roles solely through pre-screening on face-matching tests. However, even less stringent selection criteria could still provide some benefits to operational face-matching accuracy, even if only modest (Young & Noyes, 2019). By combining selection procedures with other sources of accuracy improvement, such as automated facial recognition algorithms (Section 2.4), crowd sourced decision making (Section 2.5) and support from trained forensic face examiners (Section 2.2), applied face-matching systems comprised of multiple components could be created that do meet the definition of consistent and superior face matching expertise.

---

[5] https://fullfact.org/crime/police-numbers/#:~:text=The%20number%20of%20police%20officers,Police%20and%20those%20on%20secondment).

## *2.4. Automated facial recognition algorithms*

Automated facial recognition algorithms have increased in accuracy dramatically in the last five years, due to advances in deep learning techniques and the immense quantities of facial images now readily available as training data (Noyes & Hill, 2021). State-of-the-art algorithms have shown comparable levels of performance to both face examiners and super recognisers on a challenging face matching task (Phillips et al., 2018). This section will provide an overview of the development of automated facial recognition and demonstrate why the latest state of the art algorithms could be considered 'experts' in face matching.

The use of computer programs to provide domain specific expertise is not novel. In the 1960s computer scientists began to develop programs that used artificial intelligence techniques to solve complex problems in specific domains, these programmes were known as 'expert systems'. One of the earliest expert systems, named Heuristic DENDRAL, was used to determine the molecular structure of organic compounds based on mass spectrometry data (Buchanan et al., 1969). Expert systems have since been applied to a wide range of disciplines, including searching for oil, diagnosis of disease and space exploration (Durkin, 1990). Early expert systems consisted of two components, a knowledge base and an inference engine. The knowledge base contained specialised information derived from human experts in the relevant domain. Knowledge bases typically comprised a series of 'IF' and 'THEN' statements, which are used by the system to work out a solution to the problem based on observed data via cause and effect (Durkin, 1990). The second component, the inference engine, was designed as an analogue to the 'system 2' approach of human reasoning, discussed previously. Using information from the knowledge base, the inference engine attempts to arrive at a particular conclusion by either forward chaining (using observed information to infer new information) or backward

chaining (reasoning backwards to prove if a particular hypothesis is true) (Durkin, 1990). Despite being used successfully in a number of fields, these rules based expert systems were inherently limited and could only be applied to well understood problems within narrow domains (Medsker & Turban, 1994).

Early automated facial recognition algorithms, developed in the 1960s, were similar in design to rules-based expert systems. For early algorithms, faces were defined using feature sets designed by human experts, for example using distances between facial landmarks or the size, shape and position of specific facial features (Turk & Pentland, 1991). Extracted features were then compared to a database of faces to determine if a matching face was present. These human devised features were often either highly subjective (e.g. length of ears, thickness of lips) or required the placement of specific facial landmarks to calculate distances, necessitating a human operator to extract the features within an image before submitting to the algorithm for comparison to a database. Early facial recognition algorithms required facial images taken at near identical pose and camera angle to be effective, which largely excluded them from use in real-world environments (Ballantyne et al., 1996). The feature extraction process was also time consuming and cumbersome, another major limitation in the use of such systems.

A major step change in facial recognition algorithm accuracy occurred in the 1990s, with the introduction of eigenfaces as a means to encode facial information. Rather than use the human described features of early techniques, which were limited to a particular image of a face, the eigenfaces approach attempted to characterise the variation between different facial images (Turk & Pentland, 1991). Eigenfaces are a type of eigenvector derived from principal component analysis (PCA), a dimensionality reduction technique used to reduce complex, high dimensionality data into a set of new variables. These new variables are the principal components that best explain the variance within the dataset. A facial image can

therefore be represented by the ‘best’ eigenfaces that represent the variance in appearance from the average face within a multi-dimensional space, sometimes referred to as “face space” (Turk & Pentland, 1991). The eigenfaces approach attempts to model the holistic processes humans use to recognise faces, and the technique dominated automated facial recognition algorithm development throughout the 1990s and 2000s (Masi et al., 2019). The predominant facial recognition algorithms in the 1990s and 2000s comprised of two processing stages; a feature extraction stage and distance metric learning stage (Noyes & Hill, 2021). The feature extraction stage results in a statistical representation of a face for the distance-metric phase. Unlike earlier algorithms the feature extraction stage was now automated enabling real-time operation of automated facial recognition systems. Once the features are extracted from a facial image (e.g. the eigenfaces) a model must be trained to calculate a similarity score. The similarity score indicates the similarity between two faces based on the extracted features, this is the distance metric learning stage. Distance-metric models are trained using ground truth datasets of facial images using either unsupervised machine learning (e.g. PCA) or supervised machine learning (e.g. support vector machines) image classifiers. Once trained on a dataset of faces, similarity scores can be generated for new faces, using the machine learning model. By setting a criterion score the algorithm can be used to identify matching and non-matching facial image pairs, depending on whether a similarity score between two faces exceeds the threshold of the criterion.

In the late 1990’s the US government began the first systematic testing of commercial facial recognition algorithms (Phillips et al., 1997), leading to the creation of the Facial Recognition

Vendor Test (FRVT) programme[6]. Results from the FRVT showed a steady increase in facial recognition algorithm accuracy throughout the 2000s and the early 2010s, however, algorithms were still highly constrained by pose, requiring fairly controlled front facing images. Algorithm development during this time would focus on one specific aspect of image quality, such as illumination or expression. As a result of this piecemeal approach many algorithms still performed poorly when processing uncontrolled facial images, with reports of unsatisfactory performance in real-world settings (Masi et al., 2019). Despite these limitations, some algorithms were beginning to show comparable levels of performance in face matching to humans, when matching constrained images. Three algorithms from the 2006 FRVT were shown to be comparable to the average human score when matching what the researchers considered to be 'difficult' facial images (O'Toole et al., 2008). Though varied in illumination and expression, these images were still constrained, being front facing and of fairly high resolution, and as such not representative of the types of images that might be encountered operationally in applied settings. Therefore, algorithms were recommended for use only in similarly constrained operational environments where facial images are restricted to front facing still images (Phillips & O'Toole, 2014) (e.g. passport or driving licence images as opposed to CCTV or social media images).

Up until around 2012, the techniques used for feature extraction and distance-metric learning were based on image-specific facial information rather than identity-specific information, preventing generalised face-matching performance across variable images of

[6] https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt

the same face (O'Toole et al., 2012). As a result, algorithms were accurate when matching facial images taken in similar conditions, rivalling lay human accuracy, but performance plummeted when faces varied in pose, expression, illumination or resolution (Beveridge et al., 2011). Algorithm accuracy increased drastically for unconstrained faces with the introduction of deep convolutional neural networks (DCNNs) in 2012 (Masi et al., 2019). DCNNs are considered the current state-of-the-art in automated facial recognition due to their superior performance on unconstrained images and ability to generalise across variable facial images (Hill et al., 2019). Based on the primate visual system, DCNNs encode facial images across multiple neural layers within the network. The output from preceding layers are pooled and each layer gets progressively more complex, eventually encoding highly specific facial attributes such as smiling or blue eye colour (Masi et al., 2019). DCNNs must be trained on huge datasets of labelled facial images, ideally including multiple highly variable images for each identity. Through the use of large, diverse training sets DCNNs learn to 'tell faces together', resulting in highly accurate face-matching expertise that can generalise to new and variable facial images (O'Toole et al., 2018). This is in direct contrast to human expertise in facial recognition, where variability can be learnt to 'tell together' familiar faces, but this expertise does not generalise to new, unfamiliar faces (Young & Burton, 2018).

Phillips et al. (2018) compared the performance of four state-of-the-art DCNN facial recognition algorithms against human participants, including face examiners and super recognisers on a challenging face-matching task. The top performing algorithm outperformed all bar one of the untrained student participants and many of the super recognisers and face examiners who took part in the study. Given the drastic increases in algorithm face-matching accuracy in recent years it is clear that automated approaches will have a part to play in applied face matching, not least because of the speed, efficiency and

consistency with which computers can carry out the task. However, even though state-of-the-art face-matching algorithms can outperform many humans in tests to date there are some critical factors that must be carefully considered before algorithms are implemented in high-risk applied settings. Because the features used by DCNNs are derived from many iterations of unsupervised learning on huge datasets of faces, the exact nature of DCNN architectures are poorly understood (Fong & Vedaldi, 2017). The lack of transparency in how DCNN algorithms operate is a major limitation in understanding how and when algorithms will make errors in real-world scenarios. Hill et al. (2019) recently demonstrated that the top-level layers within a DCNN algorithm create a highly structured representation of faces, with the upper most layers representing gender, pose and illumination information. Further research is required to understand how faces are represented in subsequent, more complex layers and whether this structure is consistent across different DCNN architectures.

Another consideration for the application of face-matching algorithms is that, like face examiners and super recognisers, DCNN algorithms are not a homogenous group. The accuracy of individual algorithms can vary substantially, due in large part to the size and diversity of the training database used to develop the algorithm (Noyes & Hill, 2021). Recent testing of commercial algorithms has demonstrated variations in performance for faces from different demographic groups (Bruveris et al., 2020). This has been widely reported as bias in the media, with particular attention given to varied performance on faces of different ethnicities, and is also referred to as differentials or demographic effects in the algorithm community (Grother et al., 2019b). Similarly to own race effects in human face matching, demographic differentials in algorithm performance must be clearly understood before algorithms are deployed in applied environments. Computer algorithms can be easily scaled to carry out face matching rapidly across colossal datasets. Failure to address the

risks of inaccurate performance when matching faces of different ethnic groups has major societal implications, potentially introducing widescale discrimination against certain individuals due to their ethnicity. This is particularly pertinent where facial recognition technology is used in high stakes situations, such as policing. Researchers are taking steps towards addressing differentials in performance, for example by using weighted sampling procedures during algorithm training (Bruveris et al., 2020), however the widespread adoption of the technology in applied settings without fully understanding the associated risks of algorithm bias in face matching systems is concerning.

Given the risk of error from substandard quality images in operational settings, the National Institute for Standards and Technology advised that human adjudication of algorithm results is still necessary (Grother et al., 2019a). However, due to the fact that humans are highly variable and generally poor at unfamiliar face matching there needs to be a concerted effort from the research community to quickly understand how human operators should be adjudicating face-matching algorithm results and what skills and expertise are required for such a role. The next section will look at promising research that has combined human and algorithm responses to face matching tasks, resulting in significant gains in accuracy.

## *2.5. Face-matching systems*

The majority of the face-matching discussion so far has focussed on individual sources of face-matching expertise, including face examiners, SRs, algorithms and training strategies to improve a person's face-matching accuracy. In practice, face-matching decisions are carried out within complex systems comprising multiple human components working individually or as part of small teams. With the rise in accuracy of machine learning algorithms, computer components are also becoming increasingly common within applied face matching systems (Towler, Kemp, et al., 2017). The face-matching literature to date has largely focussed on the accuracy of individual humans and algorithms and as a result the overall performance of complex systems containing multiple face matching components are not known. There is also little understanding of how the different components within the system interact and collaborate in decision making.

A recent case resulting in a wrongful arrest due to an incorrect face-matching identification provides some insights into how applied face- matching systems operate and how, in this example, errors can occur (Hill, 2020). In January 2020 an African-American man was wrongfully arrested for high value shoplifting after being confirmed as a match in a photo line-up, following a search of a CCTV still of the shoplifter from the store against a state police database of 49 million images, using a computer algorithm. The algorithm returned a ranked candidate list of high scoring images, which were then manually reviewed by a face examiner who had received at least 40 hours training in face matching[7]. The examiner

[7] https://www.michigan.gov/documents/msp/Facial_Recognition_FAQ_666807_7.pdf

concluded that a driving license image returned in the candidate list was a viable match. A second trained examiner then reviewed and confirmed the result, which was then released to investigators as an 'investigative lead', which on its own is not probable cause for arrest. In an attempt to corroborate the identification investigators showed a photo line-up, including the driving license image, to a security contractor from the store, who further confirmed the identification, despite, it would appear, having only reviewed the CCTV footage.

In this instance the initial examiner's decision was wrong and this error was not detected in the face matching system until the innocent suspect was able to confirm his alibi in a police interview room. This case highlights the need, not just to verify the expertise and accuracy of individual human face-matching operators and algorithms in applied systems but also the overall accuracy of the system, in order to develop a clearer understanding of how the individual components of the system interact as a whole.

### 2.5.1. Group decision making

Extensive early research into the effectiveness of task-orientated groups at problem solving and decision making found that interactive groups of individuals often performed poorly on tasks, whereas non-interacting groups on average performed much better (Hackman & Morris, 1975). For interacting groups, effectiveness is driven in part by ongoing interactive processes that can change throughout time and are multifaceted. These interactive processes include individual satisfaction, group cohesion, attitude changes, group size and structure (Hackman & Morris, 1975), as well as the recruitment of good collaborators and working within a conducive social context (Cantor et al., 2020. This complexity makes assessing the effectiveness of interactive groups particularly challenging in research settings and has so far been largely unexplored in applied face-matching research.

The benefits of group decision making from multiple non-interacting observers, however, are much easier to model and this approach has been demonstrated to be highly effective in face-matching research. Referred to as the wisdom of crowds (Surowiecki, 2004) or swarm intelligence (Rosenberg, 2016) this phenomenon is not unique to face matching and has been repeatedly demonstrated across a diverse range of decision-making tasks. In the early 20th century Sir Francis Galton averaged the responses of 787 individual attempts at guessing the weight of an ox at a county fair. Much to Galton's surprise the resulting crowd sourced estimate was only one-pound shy of the actual weight. Harnessing the wisdom of crowds in this manner can result in dramatic increases in decision-making accuracy, but it is important to note that the crowd is not always right. Surowiecki (2004) sets out a range of conditions that are required to effectively utilise the wisdom of the crowd. Crowds must consist of a diverse range or individuals, who are truly independent from each other and decentralised, requiring decision makers to be both specialised and free from centralised influence.

Crowd diversity is seen to be particularly important in harnessing collective wisdom. This can be diversity in terms of an individual's identity (i.e. a person's cultural experience, ethnicity, training and expertise) as well as functional diversity (i.e. how a person represents and solves problems) (Hong & Page, 2004). Studies have found that having diversity in crowds can be more beneficial than crowds based solely on measures of expertise (Hong & Page, 2004; Krause et al., 2011). This is because innovative solutions to problems will occur more quickly in large and diverse crowds due to variation in problem-solving behaviour (Cantor et al., 2020). Ideally, crowds should contain high performing diverse individuals for maximum effectiveness (Hong & Page, 2004). Given that there are significant individual differences in face-matching ability (Lander et al., 2018) and people are diverse in how they approach the task, demonstrated by individual differences in decision criteria

for matching and non-matching face pairs (Gentry & Bindemann, 2019) and variation in gaze strategies from eye-tracking data (Bobak et al., 2017), the wisdom of crowds appears to be a promising approach for improving face-matching accuracy.

### 2.5.2. Crowd effects in face matching

Aggregating scores from multiple individuals as a crowd response is one of the most effective ways of consistently improving human performance on face matching tasks, with crowds of sufficient size being on average more accurate than the top individual performers. White et al. (2013) were the first to demonstrate the wisdom of crowds on a face-matching task. Referred to as 'crowd effects', they found that by averaging responses from the normative data of the original GFMT, crowd sizes of four could be more accurate than individual top performers. A crowd size of 32 achieved near perfect performance. The study also demonstrated that the way responses are recorded and averaged impacts on the performance of the crowd. In the first part of the study individuals responded as to whether they thought the faces were a match or not match. Individual responses for a face pair were then averaged to form a crowd response. For scenarios where half the individuals in a crowd responded match and the other half responded non-match an arbitrary same decision was recorded. This led to a match bias within the data and actually impaired the performance of two-person crowds. In the second part of the study a seven-point Likert scale of similarity was used to increase diversity in responses. Averaging responses from the Likert similarity scale found improvements of 5% for crowd sizes of two and near-perfect performance for crowds of eight or more. Importantly, the benefits from crowd responses were observed for both match and non-match pairs.

The benefits of crowd effects also appear to be independent of other methods of improving accuracy in face matching tasks. Noyes (2016) found that for low resolution images,

presenting the images as a blurred rather than blocky or pixelated version improved performance on a face matching task. Combining the accuracy improvements from blurring the images with a wisdom of crowds approach provided additive gains in accuracy. This is a significant finding for applied settings, where aggregating responses of individuals into crowds could be combined with other means of improving comparison accuracy (e.g. feature-by-feature morphological comparison or using multiple images) to bring further gains in overall face comparison performance.

Optimal crowds are comprised of individuals who are both diverse problem solvers and high performers, resulting in greater gains in accuracy with smaller crowd sizes. Aggregating the responses of face examiners on a quick decision face-matching task gave gains in accuracy for crowds of two or more, which far surpassed the performance of same sized crowds of untrained controls, requiring a crowd size of four or more controls to give comparable performance to a single examiner (White, Phillips, et al., 2015). In operational settings, where resources are limited there is a trade-off between efficiency and accuracy in terms of the number of individuals working on a single face-matching task at any one time. Aggregating responses from face-matching experts reduces the crowd sizes required to achieve high performance.

Phillips et al. (2018) applied the wisdom of crowds to groups of face examiners and super recognisers on a challenging face matching task. In contrast to previous studies, examiners were supplied directly with the images and permitted to apply their agency procedures and processes. Examiner crowds of four reached ceiling performance, with an average AUC score of 1. Crowds of four untrained controls achieved an average AUC score of just over .7, with the maximum crowd size of 10 controls giving an average AUC score of .88, below that of the average individual examiner. Given that examiners were applying their standard procedures and processes it might be expected that there would be greater consistency in

their face matching decisions. It is, therefore, somewhat surprising that crowd effects gave such a drastic increase in performance for this group, as diversity is understood to be a key driver of performance in small crowds. Phillips et al. do not elaborate on the composition of different crowds or what might be causing the increases in performance. As seven individual examiners in the study are at ceiling already it may be that these high performers are the main contributors to the high accuracy of examiner crowds. However, given there is a wide range in individual examiner performance there may also be diversity in their matching decisions, leading to the prominent crowd effects. This raises questions as to how examiners develop their expertise and whether this is achieved solely through training. In this case, if examiner training is consistently delivered, it may be expected that examiners are less diverse in their face-matching strategies, as they have all received similar training. If, on the other hand, examiners are diverse in their face matching strategies this only strengthens the need for large scale examiner white box testing to better understand the nature of their expertise (see Section 2.2.3).

Phillips et al. (2018) liken the fusion of multiple examiners to the common forensic practice of verification, where two or more examines review the same case. Verification using multiple independent examiners is the recommended best practice for forensic face matching, as part of the ACE-V framework (analyse, compare, evaluate and verify) used across many forensic pattern-matching disciplines (European Network of Forensic Science Institutes, 2015, 2018). Arguably, this analogy of crowd effects and forensic verification is possibly a misinterpretation. The aim of verification, as Phillips et al. state, is to encourage consistency through a consensus of opinion, whereas the wisdom of crowds requires diverse decision makers and is actually hampered by consistency. Verification has largely arisen as a means to mitigate bias in forensic examinations (Kassin et al., 2013) rather than as means to improve accuracy specifically, though there is evidence that verification is

effective at reducing errors in fingerprint examinations (Ulery et al., 2011). It is very important that researchers do not incorrectly associate results from controlled research studies with applied techniques. Doing so risks both conflating the benefits of different approaches that may operate by distinctly different mechanisms and masking their respective limitations. Further investigation is thus required to better understand why crowd effects benefit examiner groups and how this relates to the applied practice of forensic verification.

Phillips et al. (2018) also found that the super recogniser group benefited from crowd effects. For SRs a crowd of three individuals resulted in an average AUC score of 1, meaning SR crowds reached ceiling with a smaller number of participants compared to examiner crowds, which required four participants to achieve an average AUC of 1. This is interesting given that individual SRs had a lower average score than examiners and a smaller selection pool to form crowds from (examiner N=57, SR N=13). This suggests that the face-matching strategies of SRs are highly diverse, more so than that of examiners. For applied settings this finding is also of relevance, as examiners are understood to require lengthy training to develop expertise, whereas SRs are believed to be experts at face matching innately. Therefore, recruiting a smaller crowd of three SRs would be more efficient and less costly than recruiting and training a crowd of four examiners. Further research is required to better understand how different types of human face-matching expertise can be combined and optimised in different applied settings.

A number of studies have demonstrated the benefits of crowd sourcing face-matching decisions from multiple independent observers and that, importantly, crowds of face-matching experts resulted in greater gains in accuracy with smaller numbers of individuals than crowds of non-experts (Phillips et al., 2018; Towler, White, et al., 2017; White, Phillips, et al., 2015). Of all the techniques discussed so far, the wisdom of crowds demonstrates

the strongest and most consistent benefit for improving human face-matching accuracy. With appropriate workflows and procedures, the technique could be implemented in applied face-matching systems that comprise of multiple human operators. However, it is not fully understood how applied face-matching systems operate and some may use a more collaborative approach to decision making.

### 2.5.3. Collaborative face matching

Although the wisdom of crowds is effective at improving face-matching accuracy, by keeping all observers independent within the crowd other potential benefits to face matching performance may be overlooked, such as learning and communication effects, which can occur in interacting groups (Hong & Page, 2004). There may also be applied scenarios that require a more collaborative approach to face-matching decision making.

Jeckeln et al. (2018) found no significant difference in accuracy between pairs that worked independently (non-social dyads) over pairs that worked together on a face-matching task (social dyads). Although performance was observed to be equivalent, the mechanism that underpins the decision-making process differed between social and non-social dyads. For social dyads the decision was most heavily influenced by the top performer in the pair, which was indicative of their collaborative approach. The study did not look at whether larger social and non-social crowds were equivalent in accuracy, nor if there was any learning effect for the lower performing individual within a social dyad. Dowsett & Burton (2015) did find a learning benefit from collaborative pairs where the lower performing member showed improve matching accuracy after working in a pair (see Section 2.1.6).

Combining individuals into interacting and non-interacting pairs have both been shown to improve face matching accuracy, but in different ways. In interacting pairs (or social dyads) the higher performer drives performance (Dowsett & Burton, 2015; Jeckeln et al., 2018).

However, the mechanism by which this performance gain is communicated in collaborating pairs is not known. People generally have poor insights into their own face matching ability, with low performers overestimating their own ability (Zhou & Jenkins, 2020). Feedback was not provided in either study by Jecklen et al. or Dowsett & Burton. For non-interacting pairs Jeckeln et al. (2018) found that selecting the decision of the most confident individual did not improve accuracy to the same extent as seen for collaborative social dyads. Further investigation is required to better understand how high performers improve accuracy in collaborative face-matching pairs and how this mechanism interacts with different types of face-matching expertise. For example, are social pairs equally effective for face examiners and super recognisers, and how different in ability do the two individuals need to be to increase performance and provide a training effect? Understanding these questions will help to design effective collaborative face-matching systems and reduce errors in operation face-matching tasks.

### 2.5.4. Human computer interactions

With the increasing use of automated facial recognition technology within applied face-matching systems comes an increase in human computer interaction. For example, when an individual presents their face to an e-gate at the airport for matching to their travel document, if the algorithm similarity score is below a predefined threshold that person will be sent to a human operator for verification. In a police investigation an unidentified image can be searched against a database of known facial images with the system returning a candidate list of results based on the highest algorithm similarity scores. A human operator then reviews the candidate list for any viable matches to the unidentified image. In both scenarios there is minimal interaction between the algorithm and human components in the face matching decision other than using the algorithm as a screening tool before being passed to a human operator. This piecemeal approach does not maximise the strengths

and diversity of the human and algorithmic components. Tests have shown that in some instances poor human and algorithm interaction actually increases error rates.

White, Dunn, et al. (2015) tested 24 trained facial reviewers on a candidate list review task. Using genuine Australian passport renewal images selected by a proprietary facial recognition algorithm, participants were tasked with comparing a single image to a candidate list of possible matches. The task included target-present and target-absent candidate lists and was designed to be similar to the type of face-matching task the reviewers performed in their operational duties. Participants had to compare a single probe face to eight candidate face in a series of counterbalanced target-present and target-absent trials. The facial reviewers made an error on one in every two candidate lists and were comparable in accuracy to untrained student controls, making similar proportions of misidentifications (selecting the wrong face) and misses (not selecting the correct match when present). By having the human operator as the final decision maker in this way, the accuracy of the entire face-matching system is constrained to the abilities of the operator. White, Dunn et al. also predicted that error rates in applied settings may be higher. Due to the very low prevalence of fraud in passport renewal applications, matches from automated facial recognition searches in this context are rare, causing even lower detection rates and inflating the proportion of missed matches, known as the low prevalence effect (Papesh et al., 2018).

Other studies have found an impact on human reviewer accuracy due to differences in how algorithm results are presented. Fysh & Bindemann (2018) included additional text alongside images in a two-image face-matching task that read ‘same’, ‘different’ or ‘unresolved’. In 60% of face matching trials the text was consistent with the correct response, inconsistent in 20% of trials, and unresolved in the remaining 20%. The study comprised of 65 matching identities and 5 mismatching identities in an attempt to replicate

the low prevalence rates of non-matching passports encountered at the border. Accuracy deteriorated when both unresolved and inconsistent labels were present, with the greatest decrease in accuracy occurring when labels were inconsistent. Accuracy was observed to decrease regardless of whether participants were asked to ignore or pay attention to the text labels. These findings indicate that text cues can bias human face-matching decisions in one-to-one verification tasks. Heyer et al. (2018) investigated the impact of candidate list size on facial reviewer accuracy. 99 trained facial reviewers all with at least three months face matching experience were asked to review target present and target absent candidate lists containing 10, 50 or 100 candidates. Reviewer performance showed a steady decline as candidate numbers increased. Larger candidate lists also caused a shift in bias with reviewers more likely to respond match to a face, resulting in more false alarms. Declines in performance and shifts in bias were observed for all reviewers, even those who performed better at the task. Both studies have important implications for how to design interfaces that display algorithm results to human reviewers for one-to-one verification tasks and for the review of multiple image candidate lists.

Findings from the three studies described above (Fysh & Bindemann, 2018; Heyer et al., 2018; White, Dunn, et al., 2015) demonstrate that, in some situations, having a human as the final arbiter of the face matching decision can significantly hamper the overall accuracy of the system. As facial recognition algorithms become more accurate (see Section 2.4) it is expected that the task of verifying algorithm results will only become more challenging, as the algorithm identifies more challenging matches for the human operator to review.

The various architectures of automated facial recognition algorithms are modelled on theories of human face perception and cognition (Masi et al., 2019; Toole & Roark, 2006), however, it is likely that the features and processes algorithms and humans use to match faces are very different, given their varied performance on face matching tasks (O'Toole et

al., 2008, 2012; Phillips & O'Toole, 2014). O'Toole et al. (2007) applied the wisdom of crowds using human face-matching decisions and algorithm similarity scores, which they referred to as fusion, in order to improve face-matching accuracy. This approach combines the face-matching strategies of humans and algorithms to take advantage of the diversity in the two approaches, rather than simply having the human decision override that of the algorithm. Fusing human ratings and algorithm scores improved accuracy across the board and reduced the error rate of the best performing algorithm by half. However, this study predates the use of DCNN facial recognition algorithms that now surpass the average human at some face matching tasks and the human participants were not recruited as experts in face matching. By fusing human and machine face-matching experts it may be possible to produce even greater gains in face matching performance.

Phillips et al. (2018) investigated the wisdom of crowds using three sources of face matching expertise: current state-of-the-art DCNN facial recognition algorithms, face examiners and super recognisers. Phillips et al. fused the normalised similarity scores from four DCNN algorithms with ratings from different face examiners and super recognisers. Fusing a single examiner with the similarity score of the highest performing DCNN (A2017b) gave a perfect average AUC score of 1. Fusion of a single super recogniser with A2017b also gave a huge boost in average accuracy but did not achieve ceiling. This may suggest that examiner face matching strategies are more divergent from the algorithm than that of super recognisers, however the difference in performance is marginal. As algorithm performance declined so too did the benefits of fusion, with none of the other fused algorithms surpassing the average score of an examiner or super recogniser pair. This suggests that although diversity may be playing a role in the effectiveness of human-algorithm fusion, the face-matching accuracy of both the human and algorithmic component are highly important.

### 2.5.5. Designing better face matching systems

Where an applied face-matching system has multiple humans in the face matching decision chain, research shows the wisdom of the crowds can be used to great effect, taking advantage of diversity in human face-matching expertise. Even in scenarios where face examiners are not available, or it has not been possible to recruit super recognisers, the wisdom of crowds can still improve performance and has been observed to be far more effective than short face-matching training courses. For pairs of face matchers this effect holds true regardless of whether the individuals in the pair interact or not. Group decision making provides an effective means of improving face-matching accuracy in applied systems. However, more understanding of applied face-matching systems is required to establish how group decision making can be implemented in a way that optimises performance, but does not risk introducing bias into the workflow, particularly if groups are interactive.

With the ever-increasing use of automated facial recognition technology in applied face-matching systems there is a pressing need for more research into human-machine face-matching interaction. As a priority, researchers should focus on the testing and design of effective human-machine interfaces. For example, Heyer et al. (2018) found latency in response times to be a fairly accurate prediction of face matching errors, therefore future interfaces could incorporate response time into the decision making process where operators are advised that response that have surpassed a set amount of time may not be reliable.

Fusing the scores of algorithms and human observers increases face-matching accuracy and these gains are substantial when the top performing algorithms are fused with human experts. As for the group decision making of human operators, the implementation of

human-algorithm systems requires further research and careful consideration before it can be implemented operationally, particularly as in some jurisdictions it is a legal requirement for a trained human operator to be the final adjudicator of the face matching decision[8].

Task and work analysis of applied face-matching systems will help researchers to understand the concepts, procedures and objectives of applied face matching (Moreton et al., 2019) and aid in the design of better systems in the future, incorporating elements demonstrated in the literature to improve performance, including selection criteria for high performers, recruitment and training of face examiners, crowd effects and human-algorithm fusion.

[8] https://www.jdsupra.com/legalnews/washington-state-passes-a-landmark-48587/

# *3. Thesis overview & research aims*

Chapter 1 provided an overview of unfamiliar face matching, highlighting the difficulty of the task and the extent of individual differences in the general population, followed by a discussion of the significance of face-matching errors in applied settings and the prevalence of individual differences in the performance of operational face-matching practitioners. Chapter 2 then looked to four possible sources of face-matching expertise that could benefit applied face matching; training strategies to develop face-matching expertise and three different types of face-matching experts: forensic face examiners, super recognisers and automated facial recognition algorithms. Chapter 2 concluded with a discussion of how different sources of face-matching expertise could be optimally combined in applied face-matching systems, using collective decision making and human-machine interactions. The remainder of this thesis presents a series of empirical studies on each of the four sources of expertise discussed in Chapter 2; training, superior face matchers, forensic face examiners and algorithms, concluding with a discussion of the implications of the findings for applied face matching and possible avenues for future research.

## *3.1. Study One – An international survey of face matching training*

Despite the existence of high-level training guidelines produced by the practitioner community (European Network of Forensic Science Institutes, 2018; Facial Identification Scientific Working Group, 2012b) the content, duration and delivery of face-matching training is not widely understood in the academic research community. The aim of Study One was to address this gap in the scientific literature and to better understand how different agencies train facial reviewers and face examiners, using results collected from an international survey. The survey addressed how consistent agencies are in their training methods, whether there are differences in training approaches for examiners and reviewers and the extent to which evidence-based training practices were included. These results should help researchers to better understand the diversity in training practices between different agencies, and may help explain the individual differences observed in the performance of face matching professionals and the heightened performance of face examiners at the group level.

## *3.2. Study Two – The impact of a short training course on face matching behaviour*

A recent evaluation of four short professional face matching training courses by Towler et al. (2019) showed training courses lasting three days or less lead to little, if any, immediate benefit in face-matching accuracy. However, training may impart other benefits to face-matching abilities beyond improvements in accuracy, such as helping operators to understand the limitations of face matching or reducing the likelihood of high confidence errors. For example, face examiners are much less likely to make high confidence errors on face matching tasks (Norell et al., 2015), and indeed knowing when *not* to make a decision is understood to be one of the hallmarks of forensic examiner expertise (Towler et al., 2018). Study Two evaluated a two-day short professional training course in face matching, delivered to UK police officers and staff. The study measured trainees confidence in face matching decisions on two counter-balanced trials delivered pre- and post-training. Results were compared to a control group of police personnel who did not receive training. The study also looked at the impact of training on the accuracy and confidence of high and low performances, demonstrating differences in match and non-match accuracy between high and low performers after training.

## *3.3. Study Three - Comparing perceptual skill and crowd effects for superior face matchers and face examiners*

Super face matchers and forensic face examiners have both demonstrated superior perceptual skills in quick decision face matching tasks at the group level, but also show individual differences in performance (Robertson et al., 2016; White, Phillips, et al., 2015). Study three contrasted the perceptual skills and face matching behaviours of three high performing police face matchers selected from a pool of 28 candidates, with those of three forensic face examiners, using a series of quick decision face-matching tasks. Accuracy was compared at a group level and using individual case analysis, to understand the heterogeneity of perceptual skill in individual superior matchers and examiners. The consistency of sensitivity and response bias between tasks on different days was also examined, as well as differences in confidence when the high performers made errors. Finally, the study investigated the benefits of crowd effects in boosting accuracy and reducing the prevalence of high confidence errors. The results provide insights into the benefits of selecting high performers from small candidate pools to form face-matching 'crowds', as well as an understanding of the diversity of face-matching behaviour and perceptual skill between different high performers.

## *3.4. Study Four – Combining human and algorithm expertise*

Study Four compared the performance of 138 police controls and a high performing DCNN facial recognition algorithm on two face matching tasks; one that is challenging for humans and one that is challenging for the algorithm. Human face-matching decisions and algorithm scores fusion was achieved following the technique used by Phillips et al. (2018), to create human-machine face matching pairs. The performance of human-machine pairs were evaluated on both human-challenging and algorithm-challenging images. The study demonstrated that fusion is highly effective when the algorithm is accurate, however diversity in face-matching performance between humans and the algorithm also contributed to the benefits of fusion to a lesser extent.

## *3.5. Study Five – Operational accuracy of forensic face examiners*

In applied settings face examiners often work in small teams to complete face matching tasks, using complex processes and procedures that are far removed from the quick decision face-matching tasks often used in academic research (Moreton, 2021). Study six evaluated the operational accuracy of individual face examiners and face examiner teams from 27 different agencies on a challenging face-matching task, completed using casework procedures and tools. Differences in performance between examiner teams and individuals were compared to 65 police controls at the group level and using individual case analysis. Conservatism when making errors was also investigated for the examiner groups. Finally, examiner and control decisions were combined with scores from a high performing DCNN algorithm to understand the benefits of human-algorithm fusion techniques for examiner teams and individuals.

# 4. Study One – An international survey of face-matching training

## 4.1. Introduction

One possible explanation for the varied performance of professional face-matching operators observed in the literature may be the differences in the recruitment and training of staff by different agencies. Without knowing if training is being delivered consistently by different agencies, it is difficult to identify the factors that contribute to individual differences in performance between face-matching practitioners. This study aimed to review face-matching training practices used internationally by agencies that undertake face matching, including police forces, forensic providers and immigration services, by means of an online survey. The survey asked questions about training practices for two different types of face-matching professionals: face reviewers and face examiners.

The survey questions ascertained who delivered each agency's training, the duration of the training, the topics covered, and how the training was delivered, with the objective of determining if there are differences in how training is administered between reviewers and examiners, and if there is consistency in training within the two levels. Any observed differences between training practices for face reviewers and face examiners may shed light on the discrepancies in accuracy observed between different professional groups in the literature.

The results also assessed whether training practices include elements that have been empirically evidenced to improve face-matching accuracy in the literature, including

feedback (Alenezi & Bindemann, 2013; White, Kemp, Jenkins, & Burton, 2014), facial feature comparison (Megreya & Bindemann, 2018; Towler, White, et al., 2017), working in pairs (Dowsett & Burton, 2015) and testing perceptual skill in face matching (Bobak, Dowsett, et al., 2016).

## *4.2. Methods*

### 4.2.1. Participants

The sample consisted of 24 international agencies that undertook face matching in an applied setting, including police forces, forensic providers and immigration services at the time the survey was administered. The sample included agencies from 12 different countries, shown in Figure 3. Nine of the agencies were based in Europe (38%) , eight in North America & Canada (33%), five in Australia (20%) and two in the Middle East (6%).


*Figure 3 – Word cloud of participating countries*

Participants were recruited via email using the mailing lists of practitioner working groups in face matching. Those contacted were requested to respond on behalf of their organisation or department rather than as an individual, to prevent duplicated responses from different individuals working for the same agency. Results were collected between September and December 2017. Of the agencies surveyed 15 trained face reviewers (62%) and 18 trained face examiners (75%) (*note some agencies provide face matching services at both levels).

### 4.2.2. Procedure

Participants were provided with a link to the survey via email, which was hosted on the Qualtrics platform. Upon starting the survey, participants were presented with a body of text providing some background information, confirmation of consent and the aims of the survey.

Participants were then requested to provide the country in which their agency was based and, optionally, to provide their agency name and department to prevent duplication of agencies (all data has since been anonymised). The next page required participants to provide information about the types of face matching their agency conducted; facial review and/or facial examination and what types of material they receive for comparison. The remainder of the survey focussed upon face matching training practices, specifically who delivered training, how training was delivered, the duration of training and the content of training material. The study received a favourable opinion from the ethical committee of the Open University. Results from the survey were analysed in SPSS.

## *4.3. Results and discussion*

A summary overview of the results from the survey by agency is provided in Table 4, including type of training, method of training delivery and duration of training. The subsections below provide a detailed breakdown of responses to the survey, including statistical analyses of results between the reviewer training (n=15) and examiner training (n=18) groups and a discussion of the results per subsection.

*Table 4 – Overview of training type, delivery method and duration by agency*

| Agency | Type of training | Delivery method | | | | | Duration of training |
|---|---|---|---|---|---|---|---|
| | | Online | Independent learning | Instructor driven | One-to-one mentoring | Other | |
| 1 | Reviewer | ✓ | ✓ | | | | < 1 day |
| 2 | Reviewer | | ✓ | | ✓ | | 1-6 months |
| 3 | Reviewer | | | ✓ | | | < 1 day |
| 4 | Reviewer | | | ✓ | | | 1 day |
| 5 | Reviewer | | | ✓ | | | < 1 day |
| 6 | Reviewer | ✓ | | ✓ | | | < 1 day |
| 7 | Reviewer | | ✓ | ✓ | ✓ | | 1-6 months |
| | Examiner | | ✓ | ✓ | ✓ | | 1-5 years |
| 8 | Reviewer | | | ✓ | ✓ | | 2-4 weeks |
| | Examiner | | ✓ | ✓ | ✓ | ✓ | 5+ years |
| 9 | Reviewer | ✓ | ✓ | ✓ | ✓ | | 1-6 months |
| | Examiner | ✓ | ✓ | ✓ | ✓ | | 1-6 months |
| 10 | Reviewer | | | ✓ | ✓ | | 2-4 weeks |
| | Examiner | | ✓ | ✓ | ✓ | | 1-5 years |
| 11 | Reviewer | ✓ | ✓ | ✓ | ✓ | | 1-6 months |
| | Examiner | ✓ | ✓ | ✓ | ✓ | | 1-6 months |
| 12 | Reviewer | | | ✓ | ✓ | | 1-6 months |
| | Examiner | | | ✓ | ✓ | | 1-6 months |
| 13 | Reviewer | | | ✓ | | | 2-5 days |
| | Examiner | | ✓ | ✓ | ✓ | | 1-6 months |
| 14 | Reviewer | ✓ | ✓ | | ✓ | | 1-6 months |
| | Examiner | ✓ | ✓ | | ✓ | | 1-5 years |
| 15 | Reviewer | ✓ | ✓ | | ✓ | | 6-12 months |
| | Examiner | ✓ | ✓ | ✓ | ✓ | | 1-5 years |
| 16 | Examiner | | ✓ | ✓ | ✓ | | 1-5 years |
| 17 | Examiner | | | ✓ | | | 2-4 weeks |
| 18 | Examiner | | ✓ | | ✓ | | 1-5 years |
| 19 | Examiner | | ✓ | ✓ | ✓ | | 5+ years |
| 20 | Examiner | | | ✓ | ✓ | | 6-12 months |
| 21 | Examiner | | | ✓ | | | No answer |
| 22 | Examiner | | | | ✓ | | 1-5 years |
| 23 | Examiner | | ✓ | ✓ | ✓ | ✓ | 1-5 years |
| 24 | Examiner | | | | ✓ | | 1-5 years |

### 4.3.1. Training delivery results

Of the 24 agencies surveyed, almost 80% delivered training in face matching for facial reviewers and face examiners internally. One third got training from external agencies and only a quarter procured training from commercial providers (Table 5).

*Table 5 – Source of training by agency*

| Source of Training | Agency (n=24) | |
|---|---|---|
| Internally within your agency | 79.2% | (19) |
| Externally from another agency | 33.3% | (8) |
| Supply training to other agencies | 25.0% | (6) |
| Externally from a commercial provider | 25.0% | (6) |
| Other | 12.5% | (3) |

Table 6 shows the methods of training delivery for reviewer training and examiner training. Online training was the least common delivery method for both reviewer and examiner training but was almost twice as common for reviewer training. One-to-one mentoring was used in almost 90% of examiner training but only 60% of reviewer training. Instructor driven seminars were common in both examiner and reviewer training. Independent learning featured more prominently in examiner training but in less than half of reviewer training. Differences in delivery method between reviewer and examiner training delivery methods were not found to be significant at the 95% confidence level.

*Table 6 - Delivery methods for reviewer and examiner training*

| Delivery Method | Reviewer Training (n=15) | | Examiner Training (n=18) | |
|---|---|---|---|---|
| Online training | 40.0% | (6) | 22.2% | (4) |
| Independent learning | 46.7% | (7) | 66.7% | (12) |
| Instructor driven seminars | 73.3% | (11) | 77.8% | (14) |
| One-to-one mentoring | 60.0% | (9) | 88.9% | (16) |
| Other | 0.0% | (0) | 11.1% | (2) |

### 4.3.2. Training delivery discussion

Online training was the least common method of training delivery in the sample. There were also no agencies in the sample that used only online training delivery (see Table 4). This is a promising finding as a study of the efficacy of online face matching training found that short durations of online training (less than three days) did not improve face-matching accuracy (Towler et al., 2019). Instructor driven seminars were the most common delivery method for reviewer training and second most common for examiner training, with six agencies only using instructor driven training. Woodhead et al. (1979) found no improvement in trainee comparison accuracy after a three-day instructor driven face-matching course and Towler et al. (2019) found no improvements after a one-day course and only limited and inconsistent improvements after a three-day course. These results suggest that agencies should opt against only using instructor-driven seminars to deliver training, particularly if that training is of short duration.

Almost all of the agencies providing examiner training used some form of one-to-one mentoring (88.9%). One-to-one mentoring was the second most common delivery method for reviewer training, used by 60% of agencies. Mentoring can be used as a means of transferring tacit knowledge to trainees, as often tacit knowledge is not formally recorded and is instead based on the experiences of more senior colleagues (Mayfield, 2010). Mentoring has been found to be highly effective in the training of new teachers (Langdon, 2014) and in transferring knowledge and experience in corporate business (Sosik et al., 2004). Despite being the most common delivery method for face examiner training in this survey the author is not aware of any study to date that has evaluated the effectiveness of mentorship for improving face matching accuracy. Dowsett & Burton, (2015) found that when novices worked in pairs on a face matching task the higher performing individuals in the pair provided a learning effect for lower performing individuals, which transferred to a

second face matching task completed individually. But this design is unlikely to be representative of a professional face-matching mentorship programme. Given the widespread use of one-to-one mentoring in face-matching training reported in this survey, future research should look at how mentoring is being delivered by face-matching agencies and establish the effectiveness of this approach in improving face-matching accuracy.

Studies to date have only looked at one type of delivery method for face-matching training over short durations (Towler et al., 2019; Woodhead et al., 1979). It is important to note that the majority of agencies surveyed used multiple training delivery methods for both reviewers and examiners (see Table 4), but there was little consistency in the delivery methods used between different agencies. When deciding how to deliver training, face-matching trainers should consider the research that has already been conducted on online and short instructor-driven delivery methods. There also needs to be further work with researchers to establish what combinations of delivery methods are most effective. This collaborative approach will drive consistency in training delivery that is much needed based on the findings of the current survey.

#### 4.3.3. Training duration results

Table 7 presents a comparison of survey results for training duration between the reviewer and examiner training groups. There are clear differences in training duration between the groups. Examiner training was generally of longer duration, with 66.7% of agencies delivering training that lasted a year or more in duration and only one agency (5.6% of the sample) delivered examiner training that was less than a month in duration. Conversely, of the agencies surveyed, no reviewer training exceeded 12 months in duration and 40% of agency reviewer training was five days or less. A Mann-Whitney U test revealed that the distribution of training duration was significantly different between examiner and reviewer

training groups ($U$ = 20.5, p = .001), with reviewer training tending to be shorter than examiner training.

*Table 7 – Duration of reviewer and examiner training*

| Duration | Reviewer Training (n=15) | | Examiner Training (n=18) | |
|---|---|---|---|---|
| Less than 1 day | 26.7% | (4) | 0.0% | (0) |
| 1 day | 6.7% | (1) | 0.0% | (0) |
| 2 to 5 days | 6.7% | (1) | 0.0% | (0) |
| 2 to 4 weeks | 13.3% | (2) | 5.6% | (1) |
| 1 to 6 months | 40.0% | (6) | 16.7% | (3) |
| 6 to 12 months | 6.7% | (1) | 5.6% | (1) |
| 1 to 5 years | 0.0% | (0) | 55.6% | (10) |
| 5 + years | 0.0% | (0) | 11.1% | (2) |

Figure 4 shows the distributions of training duration for both examiner and reviewer training. Examiner training is clearly skewed towards longer durations (one year and greater). However, there is a wide range in training duration from two to four weeks to five plus years. Reviewer training durations show a bimodal distribution with 40% of agencies providing one to six months of training and 26.7% of agencies providing less than one day of training. As for examiner training there was a very wide range in reviewer training duration.

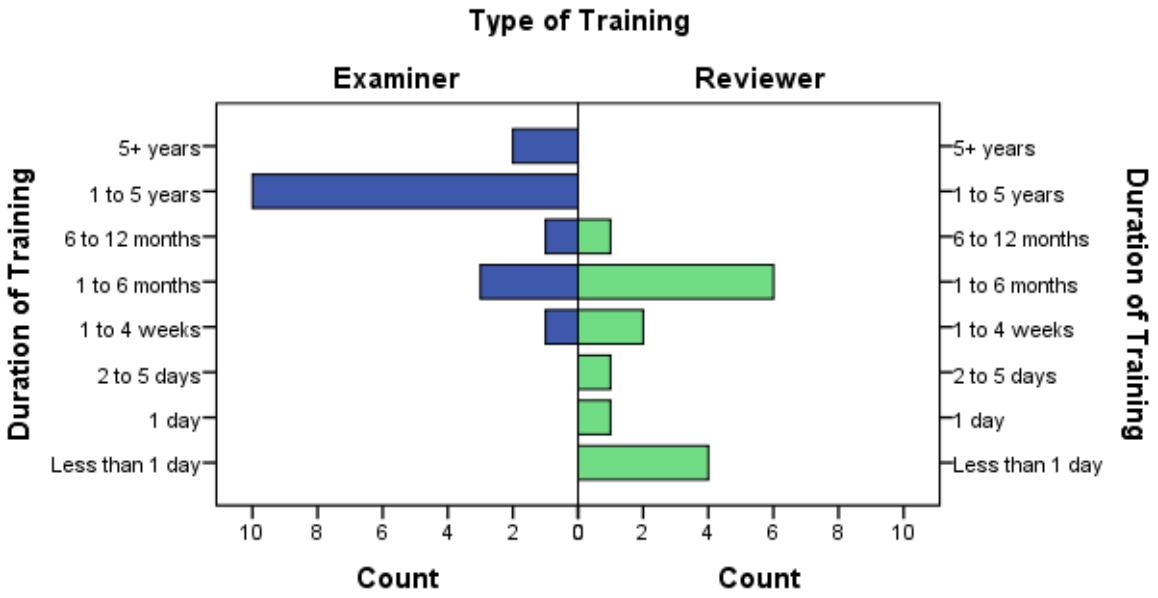


*Figure 4 – Frequency distributions for durations of examiner and reviewer training*

### 4.3.4. Training duration discussion

It is clear that as a group examiners receive much longer durations of training than most reviewers and this may offer a possible explanation for the greater accuracy of trained examiners over trained reviewers observed in research (Phillips et al., 2018; White, Dunn, et al., 2015; White, Phillips, et al., 2015). However, examiners have shown significant individual variability in face matching on both quick decision tests and more challenging comparisons that resemble operational imagery (Norell et al., 2015; Phillips et al., 2018; White, Phillips, et al., 2015). Like the wide range in training delivery methods, training duration is comparably inconsistent across different agencies. For examiners, the majority of agencies provided training for one to five years, whereas the second highest count for training duration was one to six months and the shortest durations of examiner training demonstrated from this survey is only two to four weeks. Given these substantial differences in time it can be assumed that training practices for examiners are very different between agencies. If training does have a significant bearing on examiner accuracy then the variability in training duration may contribute to the individual differences between examiners observed in the literature.

Individual facial reviewer accuracy also varies substantially in the literature, with some reviewers performing at examiner group levels (Phillips et al., 2018) and others at the level of untrained novices or even chance (White, Kemp, Jenkins, Matheson, et al., 2014). The diversity in training durations for reviewers is equally, if not more, concerning than that of examiners. 40% of agencies provided one to six months of training but another 40% provided reviewer training that was five days or less and 26% of agencies provided less than one days training. This is an alarming finding given that studies have shown limited or no improvements in accuracy from three-day training courses and no improvements from training courses that are one day or less (Towler et al., 2019; Woodhead et al., 1979). It

can only be assumed that reviewers being trained for one to six months are receiving a very different training experience to those who are trained for less than a day.

Although job-specific requirements will likely vary between agencies, the tasks and responsibilities of face-matching personnel from different agencies should be broadly comparable. That different agencies are so inconsistent in the durations of training for both examiners and reviewers is a critical issue that should be addressed as a priority. Communication and collaborative working between agencies should be the first step, which could be facilitated by practitioner working groups such as FISWG and ENFSI. Interestingly, surveyed agencies that provide training to both reviewers and examiners have longer durations of reviewer training than those that only train reviewers (see Table 1). This suggests there may be some overlap between reviewer and examiner training practices for agencies that do both.

Current research has only addressed very short durations of training (maximum three days), but the results of the current survey show that some training lasts for far longer than this, suggesting that further research should look at the effectiveness of longer-term training. When testing examiner and reviewer accuracy on face matching tasks researchers should also establish the extent of training received by participants to determine if there is a relationship between training duration and accuracy.

#### 4.3.5. Training topic results

The survey asked respondents whether training included training topics recommended by international best practice documents (European Network of Forensic Science Institutes, 2018; Facial Identification Scientific Working Group, 2010). Table 8 shows how many agencies covered each general topic. There was an overall trend for examiner training to cover more of the recommended topics than reviewer training. The only topic to be covered

by all agencies was methods of comparison and this was for examiner training. None of the differences between examiner and reviewer training were significant at the 95% confidence level.

*Table 8 – Training topics covered by examiner and reviewer training*

| Training Topics | Reviewer training (n=15) | Examiner training (n=18) |
|---|---|---|
| Anatomy | 66.7% (10) | 94.4% (17) |
| Image science | 66.7% (10) | 77.8% (14) |
| Image processing | 66.7% (10) | 83.3% (15) |
| Comparison methods | 80.0% (12) | 100% (18) |

The following sections breakdown the four recommended training topics by subtopic based on best practice documentation, with statistical analyses of results between examiner training and reviewer training using odds ratios and Fisher's exact test of significance at the 95% confidence level.

#### *4.3.5.1. Anatomy training subtopic results*

The examiner training group covered a high proportion of subtopics in facial anatomy, ranging from 94% of agencies covering ‘face shape’, ‘features of the skin’ and ‘creases and lines’ to 72% covering ‘juvenile development’ and ‘alterations to the face’. Anatomy training was covered in less detail by reviewer training, ranging from 66.7% of agencies features of the ‘nose’, ‘mouth’, ‘features of the skin’, ‘creases and lines’ and ‘facial expression’ to only 27% of agencies providing training in ‘juvenile development’ (Table 9).

*Table 9 – Anatomy training subtopics covered by reviewer and examiner training*

| Anatomy Subtopics | Reviewer Training (n=15) | | Examiner Training (n=18) | |
|---|---|---|---|---|
| Face shape | 60.0% | (9) | 94.4% | (17) |
| Eyes | 60.0% | (9) | 83.3% | (15) |
| Ears | 60.0% | (9) | 88.9% | (16) |
| Nose | 66.7% | (10) | 88.9% | (16) |
| Mouth | 66.7% | (10) | 88.9% | (16) |
| Chin and jaw | 60.0% | (9) | 88.9% | (16) |
| Features of the skin (e.g. scars/marks) | 66.7% | (10) | 94.4% | (17) |
| Bones of the skull | 46.7% | (7) | 77.8% | (14) |
| Muscles of the face* | 33.3% | (5) | 77.8% | (14) |
| Creases and lines | 66.7% | (10) | 94.4% | (17) |
| Facial Expression | 66.7% | (10) | 83.3% | (15) |
| Effects of ageing* | 46.7% | (7) | 88.9% | (16) |
| Juvenile development* | 26.7% | (4) | 72.2% | (13) |
| Permanence of features* | 40.0% | (6) | 88.9% | (16) |
| Alterations to the face (e.g. piercings) | 53.3% | (8) | 72.2% | (13) |

Statistically significant differences were observed between reviewer and examiner training for the following anatomy subtopics; ‘muscles of the face’ (OR = 9.6, 95% CI: 1.9, 47.4, $p$ = .005, Fisher's exact test), ‘effects of ageing’ (OR = 9.1, 95% CI: 1.5, 54.5, $p$ = .020, Fisher's exact test), ‘juvenile development’ (OR = 7.1, 95% CI: 1.5, 33.3, $p$ = .015, Fisher's exact test) and ‘permanence of features’ (OR = 9.1, 95% CI: 1.5, 54.5, $p$ = .020, Fisher's exact test).

All anatomy subtopics were more frequently covered in examiner training than in reviewer training, however most of these differences were not statistically significant. The lack of significance observed for certain differences between reviewer training and examiner training may be attributable to the small size of the sample in the study. For example, although it is more than eight times more likely for examiner training to cover both ‘face shape’ and ‘features of the skin’ (e.g. scars and marks) than the reviewer training group, this was not found to be significant at the 95% confidence level (OR = 8.5, 95% CI: 0.87, 83.5, $p$ = .070, Fisher's exact test). Given the large confidence interval observed for most

results, it is likely that the small sample may be of insufficient size to reject the null hypothesis even when a large effect is present (i.e. a propagation of type II errors). The lack of statistical power is likely to also be an issue for other training subtopics, not just in facial anatomy. Although only a minority of results are significant at the 95% confidence level there is a clear trend for examiner training covering proportionally more anatomy topics than reviewer training.

#### *4.3.5.2. Image science and processing training subtopic results*

Table 10 shows the proportion of agencies that covered image science subtopics broken down by training type (reviewer or examiner), Table 11 shows these results for image processing subtopics.

*Table 10 – Image science subtopics covered by reviewer and examiner training*

| Image Science Subtopics | Reviewer Training (n=15) | | Examiner Training (n=18) | |
|---|---|---|---|---|
| Properties of visible light | 40.0% | (6) | 66.7% | (12) |
| Properties of non-visible wavelengths | 33.3% | (5) | 55.6% | (10) |
| Image capture and camera sensors | 33.3% | (5) | 61.1% | (11) |
| Impact of lighting and camera exposure | 66.7% | (10) | 77.8% | (14) |
| Geometric distortions | 60.0% | (9) | 77.8% | (14) |
| Aspect ratio distortion | 53.3% | (8) | 77.8% | (14) |
| Pixel resolution | 66.7% | (10) | 77.8% | (14) |
| Image compression | 66.7% | (10) | 77.8% | (14) |
| Video compression | 33.3% | (5) | 66.7% | (12) |

*Table 11 – Image processing subtopics covered by reviewer and examiner training*

| Image Processing Subtopics | Reviewer Training (n=15) | | Examiner Training (n=18) | |
|---|---|---|---|---|
| Brightness and contrast adjustments | 73.3% | (11) | 83.3% | (15) |
| Rotations and cropping | 66.7% | (10) | 83.3% | (15) |
| Sharpening and blurring | 53.3% | (8) | 77.8% | (14) |
| Scaling | 60.0% | (9) | 83.3% | (15) |
| Colour channel separation | 46.7% | (7) | 66.7% | (12) |
| Effects on facial appearance | 60.0% | (9) | 83.3% | (15) |

All subtopics in image processing and image science were covered by more examiner training agencies than reviewer training agencies, however none of these differences were found to be statistically significant at the 95% confidence level, which may be due to a lack of statistical power as mentioned above. In general, fewer agencies covered image science subtopics than image processing subtopics for both reviewer and examiner training.

#### *4.3.5.3. Comparison method subtopics results*

The proportions of reviewer and examiner training that covered comparison method subtopics are shown in Table 12. Almost all subtopics follow the previous pattern of being covered in a higher proportion of examiner training than reviewer training. 'Use of automated facial recognitions systems' does not follow this trend and is covered by a slightly higher proportion of reviewer training agencies than examiners (73.3% and 61.1% respectively), however this difference was not statistically significant at the 95% confidence level.

*Table 12 – Comparison method subtopics covered by reviewer and examiner training*

| Comparison Method Subtopics | Reviewer Training (n=15) | | Examiner Training (n=18) | |
|---|---|---|---|---|
| Instruction in the ACE-V framework* | 53.3% | 8 | 100.0% | 18 |
| Instruction in holistic comparison | 66.7% | 10 | 66.7% | 12 |
| Limitations of holistic comparison | 66.7% | 10 | 83.3% | 15 |
| Instruction in morphological comparison | 80.0% | 12 | 100.0% | 18 |
| Limitations of morphological comparison* | 53.3% | 8 | 88.9% | 16 |
| Instruction in facial feature classification | 40.0% | 6 | 55.6% | 10 |
| Limitations of facial feature classification | 53.3% | 8 | 83.3% | 15 |
| Instruction in photo anthropometry | 26.7% | 4 | 50.0% | 9 |
| Limitations of photo anthropometry* | 46.7% | 7 | 88.9% | 16 |
| Instruction in superimposition | 20.0% | 3 | 50.0% | 9 |
| Limitations of superimposition* | 46.7% | 7 | 88.9% | 16 |
| Use of automated facial recognition | 73.3% | 11 | 61.1% | 11 |
| Human facial recognition | 73.3% | 11 | 77.8% | 14 |
| Cognitive bias | 66.7% | 10 | 83.3% | 15 |
| Own-race effects | 53.3% | 8 | 61.1% | 11 |
| Evaluating comparison findings | 60.0% | 9 | 83.3% | 15 |
| Peer-review and independent verification | 66.7% | 10 | 83.3% | 15 |

Statistically significant differences were observed between reviewer and examiner training for the following methods of comparison subtopics; ‘ACE-V framework’ (OR = 15.7, 95% CI: 1.6, 150.1, $p$ = .011, Fisher's exact test), ‘limitations of morphological comparison’ (OR = 7, 95% CI: 1.2, 41.8, $p$ = .047, Fisher's exact test), ‘limitations of photo anthropometry’ (OR = 9.1, 95% CI: 1.5, 54.5, $p$ = .020, Fisher's exact test) and ‘limitations of superimposition’ (OR = 9.1, 95% CI: 1.5, 54.5, $p$ = .020, Fisher's exact test). ‘Instruction in the ACE-V framework’ and ‘instruction in morphological comparison are the only subtopics from the survey to be covered by all 100% of examiner training agencies. ACE-V is an acronym for analyse, compare, evaluate-verify a framework used in several forensic comparison disciplines, such as fingerprint comparison (European Network of Forensic Science Institutes, 2015).

#### 4.3.6. Training topics discussion

All training subtopics were covered by a higher proportion of examiner training agencies than reviewer agencies, with the exception of ‘instruction in the use of automated facial recognition systems’. Given the longer durations of training for examiners it is not surprising that more subtopics are included in their training, and as facial reviewers are more likely to use automated facial recognition systems, it is also unsurprising that this topic is more commonly covered in reviewer rather than examiner training.

For anatomy training there were statistically significant differences observed for four subtopics (‘muscles of the face’, ‘effects of ageing’, ‘juvenile development’ and ‘permanence of features’). These four topics appear to be the more detailed and complex aspects of facial anatomy, supporting the notion that examiner training is more in-depth than reviewer training. For image science and image processing, examiner training agencies covered

proportionally more subtopics, although no differences were found to be statistically significant.

Anatomy and imaging topics were covered by a high proportion of training agencies (see Table 5) as recommended by best practice guidelines, and are explored in more detail for examiner training. However, at present research demonstrates that knowledge of these topics does not improve face matching accuracy (Towler, 2016). Therefore, it is unlikely that more in-depth training in anatomy and imaging is responsible for the enhanced accuracy of examiners, though such training may provide useful knowledge for explaining findings and observations, particularly in reports and during testimony in court. Greater inclusion of image science topics in examiner training may be a contributing factor to why examiners are more cautious than novices when comparing low quality images (Norell et al., 2015).

For the methods of comparison subtopics, four statistically significant differences were observed between reviewer and examiner training agencies; 'instructions in the ACE-V framework', 'limitations of morphological comparison', 'limitations of photo anthropometry and 'limitations of superimposition'. These differences indicate that examiner training may be more method orientated than reviewer training with a greater emphasis on the associated limitations of different comparison methods. This knowledge may also be contributing to why examiners are generally more cautious when making comparisons than novices.

It is very concerning indeed that half of examiner training agencies were providing instruction in the use of facial feature classification, photo anthropometry and superimposition, despite these methods being demonstrated to be unreliable in the literature (Kleinberg & Vanezis, 2007; Moreton & Morley, 2011; Ritz-Timme et al., 2011; Strathie et al., 2012; Strathie & McNeill, 2016; Towler et al., 2014) and best practice

guidance recommending the methods not be used in comparison (European Network of Forensic Science Institutes, 2018; Facial Identification Scientific Working Group, 2019a). It is highly inadvisable for agencies to instruct trainees in methods that have been repeatably demonstrated to be unreliable.

Methods of comparison also included the only two subtopics from the survey covered by 100% of agencies, and this was only for examiner training. These subtopics were 'instruction in the ACE-V framework' and 'instruction in morphological comparison'. ACE-V is a sequential process used in forensic comparison disciplines whereby the suspect material (e.g. CCTV video) is first analysed in isolation to determine its suitability for comparison, then compared to the reference material (e.g. a custody mugshot) and the findings from the comparison are evaluated to determine the weight of evidence. Then the process is independently verified by another examiner who repeats the process (European Network of Forensic Science Institutes, 2018). That this subtopic is covered by 100% of examiner training and only half (53%) of reviewer training supports that examiners are more likely to conduct rigorous and detailed face-matching examinations whereas reviewers often work in high throughput environments where time limits are more constrained (Facial Identification Scientific Working Group, 2010).

'Instruction in morphological comparison' was also covered by 100% of examiner training agencies. Morphological comparison refers to the process of comparing faces on a feature-by-feature basis rather than holistically (Facial Identification Scientific Working Group, 2019a). Morphological comparison has been demonstrated to improve the accuracy of novices on face matching tasks when the images are a match (Megreya & Bindemann, 2018; Towler, White, et al., 2017). The fact that all examiner training reviewed as part of this survey included both instruction in ACE-V and instruction in morphological comparison indicates that forensic face examiners are trained to compare faces in a qualitatively

different way to novices (Towler, White, et al., 2017). By being trained in comparing faces in a sequential series of processes on a feature-by-feature basis, examiners may be learning to override the intuitive, quick decision comparisons used by untrained novices and instead learn to make comparisons in a more systematic but resource intensive and time-consuming way (Towler et al. 2021).

### 4.3.7. Evidence-based training strategies results

The previous sections from the survey focussed on training topics recommended by face-matching practitioner working groups. This section of the survey asked respondents whether their face-matching training programmes included any evidence-based exercises, which have been demonstrated to improve face-matching ability in the literature. Table 13 shows the evidence-based approaches to training included in the survey with supporting references.

*Table 13 – Evidence-based approaches to improving face matching accuracy*

| Evidence-based approach | Reference |
|---|---|
| 1. Face matching exercises<br>2. Feedback on comparison responses | White et al., 2014. Feedback training for facial image comparison |
| 3. Comparison of specific facial features | Megreya & Bindemann, 2018. Feature instructions improve face-matching accuracy |
| 4. Working on tasks in pairs or groups | Dowsett & Burton, 2015. Unfamiliar face matching: Pairs outperform individuals and provide a route to training |
| 5. Face matching using a facial feature list | Towler, White, et al., 2017. Evaluating the feature comparison strategy for forensic face identification |
| 6. Enhancement of images for comparison (e.g. blurring pixelated images) | Noyes, 2016. Face recognition in challenging situations |
| 7. Testing perceptual skill prior to training<br>8. Testing perceptual skill after training | Bobak, Dowsett, & Bate, 2016. Solving the border problem: Evidence of enhanced face matching in individuals with extraordinary face recognition skills |

Table 14 shows the number of agencies that included evidence-based approaches in reviewer and examiner training. The majority of evidence-based approaches followed the trend of other topics in the survey by being more widely adopted in examiner training. Testing face matching ability before and after training, however, were both more common in reviewer training courses. Evidence-based approaches have a very wide range in adoption from 27% for working in pairs or groups on face matching tasks, up to 94% for inclusion of face matching exercises. The only significant difference observed between reviewer and examiner training was in the use of facial feature checklists in face matching (OR = 7, 95% CI: 1.2, 41.8, $p = 0.047$, Fisher's exact test).

*Table 14 – Research-based approaches used in reviewer and examiner training*

| Research-based approach | Reviewer Training (n=15) | | Examiner Training (n=18) | |
|---|---|---|---|---|
| Face matching exercises | 93.3% | (14) | 94.4% | (17) |
| Feedback on comparison responses | 66.7% | (10) | 88.9% | (16) |
| Comparison of specific facial features | 60.0% | (9) | 83.3% | (15) |
| Working on tasks in pairs or groups | 26.7% | (4) | 33.3% | (6) |
| Face matching using a facial feature list* | 46.7% | (7) | 88.9% | (16) |
| Enhancement of images for comparison | 60.0% | (9) | 83.3% | (15) |
| Testing perceptual skill prior to training | 40.0% | (6) | 38.9% | (7) |
| Testing perceptual skill after training | 60.0% | (9) | 55.6% | (10) |

### 4.3.8. Evidence-based training strategies discussion

Individual differences in face matching have been consistently demonstrated in the face matching literature, for many years (see Lander et al., 2018). However, only 40% of agencies tested personnel prior to enrolment in training. Surprisingly, more agencies tested face matching ability after training (up to 60%). Researchers have advocated pre-screening of face matching accuracy as a means to identify high performers for recruitment into operational face matching roles (Bobak, Dowsett, et al., 2016; White et al., 2021). This is of particular pertinence given that, for facial reviewers, professional experience has been

demonstrated to bear no correlation to face matching accuracy (White, Kemp, Jenkins, Matheson, et al., 2014) and in the majority of studies reviewers perform at comparable levels to controls (White et al., 2021).

Almost all agencies included face matching exercises of some form in their training, barring one reviewer and one examiner training programme. Of those that do, not all provided feedback on responses, despite empirical evidence that feedback on face-matching decisions can improve accuracy (White, Kemp, Jenkins, & Burton, 2014). Feedback on tasks has been demonstrated as a key contributor to developing perceptual expertise (Edmond et al., 2017).

Training approaches that encourage the use of facial features matching stratgies were more common in examiner training and significantly more examiner training programmes included the use of facial feature checklists. Facial feature comparison (Megreya & Bindemann, 2018) and feature checklists (Towler, White, et al., 2017) have both been demonstrated to improve accuracy for matching face pairs. The greater use of facial feature comparison and feature checklists in training may be a contributor to the enhanced accuracy of examiners at the group level.

The variable adoption of evidence-based approaches in face-matching training, with some of the lowest adoption rates observed in the entire survey, suggests that there is a disconnect between face-matching research and professional training practices. Therefore, greater communication and collaboration is required between face-matching researchers and practitioner working groups that develop training guidelines, to ensure more widespread adoption of evidence-based approaches that have been empirically proven to improve face-matching accuracy.

## *4.4. General Discussion*

To date there has been no longitudinal study of face matching training to validate the effectiveness of longer training programmes. Only short professional training courses of up to three days have been empirically tested (Towler et al., 2019) and have been shown to have little, if any impact on improving face-matching accuracy. For the agencies that took part in the survey all examiner training and almost half of reviewer training is at least one week in duration, with most reviewers trained for one to six months and most examiners for one to five years. Searston & Tangen (2017b) tracked the performance of 24 fingerprint trainees over the period of one year, with tests measuring four aspects of fingerprint expertise delivered every three months. Using a composite score of the four measures, they found the most significant gains in expertise to occur in the first three months with gains steadily plateauing after this point. Future research should investigate whether face-matching expertise shows a similar emergence over time and determine if longer durations of training could be the source of enhanced examiner performance.

Searston & Tangen (2017b) also found that the initial perceptual expertise of trainees at fingerprint matching prior to training was a reliable predictor of ongoing performance throughout the training programme. This means that for fingerprint expertise, pre-screening of applicants appears effective in identifying high performers who will continue to develop enhanced expertise through training. Pre-screening using face-matching tests has been widely advocated for identifying high performing individuals for applied roles (Bobak, Dowsett, et al., 2016; Ramon et al., 2019a; White, Dunn, et al., 2015), but to date no study has investigated the interaction between pe-screening and long term training in applied face matching. According to the survey only 40% of agencies were pre-screening the face-matching abilities of reviewers and examiners prior to training. If the development of

fingerprint-matching expertise is analogous to face-matching expertise then pre-screening may be useful tool in evaluating the effectiveness of longer term training and identifying the top performers for subsequent enrolment into training.

Studies of face-matching training have found short online and instructor driven courses to be largely ineffective at improving face matching ability. However, from the survey results it appears agencies are using a range of different training delivery techniques for reviewer and examiner training, including independent learning and one-to-one mentoring. In-house mentoring is seen to be an important part of professional development across the forensic sciences (Ashcroft et al., 2004) and practitioner working groups in face matching recommend that facial reviewers and examiners be assigned a workplace mentor during their training and development (European Network of Forensic Science Institutes, 2018; Facial Identification Scientific Working Group, 2019d). As yet no study has evaluated the effectiveness of workplace mentoring in developing face matching expertise. As recommended by Towler et al. (2019), mentoring in face matching should also be investigated as a part of the evaluation of longitudinal face matching training.

Researchers also advocate the use of evidence-based training practices to develop perceptual expertise (Searston & Tangen, 2017b; Towler et al., 2019). The results from the survey demonstrate that the use of evidence-based training practices are relatively low for both examiner and reviewer training. Whilst training guidelines for face-matching practitioners do exist (e.g. Facial Identification Scientific Working Group, 2010), these guidelines are prescriptive on the topics that a training course should include but are lacking details on how to design evidence-based training courses that develop and measure emerging perceptual expertise over time. In this regard, there needs to be greater collaboration between the practitioner community and relevant researchers in human-

performance testing, cognitive science and psychology to design empirically-derived face-matching training that is proven to develop expertise.

The survey results have revealed a general trend where examiners are trained for longer than reviewers, via a more diverse range of delivery methods and trained using a wider range of topics and research-based approaches. These observations may provide an explanation for the enhanced face-matching accuracy of examiners at the group level. As well as there being overall differences in duration, delivery techniques and topics between reviewer and examiner training at the group level, there are also substantial individual differences in training practices within each group. Perhaps the most notable difference within both reviewer and examiner training is duration. Reviewer training ranged from less than a day to up to 12 months. 40% of agencies run reviewer training for five days or less, this is particularly alarming given that training courses of three days and less have been demonstrated to provide no consistent improvement in face matching accuracy (Towler et al., 2019). In the upper range of reviewer training duration, 40% of agencies provided one to six months of training. Likewise, for examiners training durations ranged from one to four weeks to up to five plus years. Differences in training practices were also observed for delivery methods, topics and inclusion of research-based exercises. This diversity in results indicates a lack of standardisation in training practices between different agencies and may be a contributing factor towards the individual differences in reviewer and examiner accuracy observed in the literature (see White et al., 2021).

In addition to longitudinal training studies and the development of evidence-based training practices, the practitioner and research communities should also move towards white box testing of examiners and reviewers. In white box testing participants disclose information about the processes and procedures they used to complete the test. White box testing has been used to better understand the expertise of forensic fingerprint examiners (e.g. Ulery

et al., 2015). In a white box testing scenario agencies could report how their personnel are selected and trained. By disclosing the extent of training that reviewers and examiners have received in a white box face-matching test, relationships between different training approaches and face-matching performance may be found. In this manner it may be possible to identify agencies with effective training regimes and use these as a model for the wider face-matching community, which could in turn reduce the wide range in individual differences within these practitioner communities and mitigate against high risk errors in applied face-matching scenarios.

# 5. Study two – The impact of a short training course on face-matching behaviour

## 5.1. Introduction

Chapter 4 revealed that some agencies rely on short training courses of five days or less to train face-matching personnel, who are referred to as facial reviewers by the practitioner community (Facial Identification Scientific Working Group, 2019a). Recent research by Towler et al. (2019) has demonstrated that short training courses do not consistently or reliably improve face matching accuracy. Towler et al. evaluated four short professional face matching training courses, ranging from one hour to three days in duration, using pre- and post-training face-matching tests. For three of the four courses, no difference in trainee face-matching accuracy was observed after training. For the three day training course some minor improvements were observed but only for certain face stimuli and not for uncontrolled face images that most resembled those encountered operationally. These findings, and similar results from a much earlier study by Woodhead et al. (1979), strongly suggest that short training courses alone do not provide immediate improvements in face-matching accuracy.

However, training may impart other benefits beyond improvements in accuracy, such as an improved understanding of the strengths and limitations of a technique. A hallmark of forensic face examiner expertise is that they are less likely than novices to make high-confidence face-matching errors (Norell et al., 2015), and this expertise is believed to be derived from training and experience (Towler et al. 2021). Conversely, untrained super recognisers, with comparable levels of accuracy on a face-matching task to a group of

trained examiners, made a notably higher proportion of extremely confident errors (Phillips et al., 2018). If an incorrect face-matching decision is given undue confidence by a face matching operator it is possible that this could lead to further, incorrect action, thus propagating and compounding the impact of the error. Ensuring that face-matching decisions are made with an appropriate and well-calibrated level of confidence is important in applied settings where subsequent actions, with potentially life-changing consequences, could be made based on that decision, such as wrongful arrest or conviction.

The role of confidence has been widely explored in face recognition tasks, such as eyewitness identification (Iida et al., 2020; Wixted et al., 2016; Wixted & Wells, 2017), however, there has been little investigation of the confidence-accuracy relationship in face matching. Stephens et al. (2017) found that on face-matching tasks consisting of equal numbers of matching and non-matching trials counterbalanced for difficulty, the confidence ratings of novice face matchers were a reliable predictor of overall accuracy. However, some asymmetry was observed in confidence-accuracy between matching and non-matching trials, with participants being generally more confident on matching trials. As the proportion of matching trials increased so too did over confidence in the decisions, caused by a shift in response bias. Similarly, Fysh & Bindemann (2017c) demonstrated the emergence of a liberal response bias in a prolonged face-matching task, with participants developing a greater propensity to respond 'match' over time resulting in reduced accuracy on non-matching trials. In an earlier study, enforced breaks and desk switching did not alleviate a decline in non-matching performance (Alenezi et al., 2015), whereas feedback on trials did (Alenezi & Bindemann, 2013). White, Kemp, Jenkins, & Burton (2014) found that feedback on face matching trials resulted in participants being more confident when correct, improving their confidence calibration. However, improvements in accuracy from feedback training in this study were confined largely to participants who demonstrated

initially poor face-matching performance. Gentry & Bindemann (2019) used example label matching and non-matching face pairs as a training aid, which also increased the accuracy of low performers but only for the face stimuli from the same database as the examples. They believed the examples helped to stabilise the inconsistent decision criterion of low performers. Ideally, an effective face-matching training course would result in a better calibrated confidence-accuracy relationship and reduce bias in responses (e.g. Gentry & Bindemann, 2019; White, Kemp, Jenkins, & Burton, 2014). Alternatively, training could introduce a response bias, if the training teaches strategies that increase the use of match over non-match responses or vice-versa. Training can also cause over confidence, with the confidence of beginners increasing despite no increase in actual ability (Sanchez & Dunning, 2018). Both of these outcomes would be undesirable in applied settings.

The aims of this study were to explore whether a two-day face-matching training course has any immediate impact on face matching behaviour beyond changes in accuracy. These include changes in the use of high confidence decisions, response bias (i.e. propensity to respond match or non-match) and whether training affects high and low performers differently.

## *5.2. Method*

### 5.2.1. Participants

27 police trainees (14 female, age range 25 to 64 years, median age category 35 to 44 years) participated in the study. Trainee participants had not received any prior training in face matching but did use face matching in their day-to-day duties.

Initially 31 police controls completed both parts of the study, however three control participants were removed as outliers due to one participant performing substantially below chance on matching face pairs and two participants displaying a major change in response bias between the two trials. This resulted in 28 police controls (7 female, age range 18 to 64 years, median age category 35-44 years). Control participants had not received any training in face matching and predominantly worked within the field of digital forensics.

### 5.2.2. Materials

In order to test for any changes in face-matching accuracy or behaviour post training participants in the trainee and control groups completed two counterbalanced face-matching trials (trial A and trial B) similar to those used by Towler et al. (2019). Trials were counterbalanced in terms of face pair difficultly and the number of matching and non-matching pairs. Trial A consisted of 107 face image pairs and trial B consisted of 108 face image pairs. Images for the trials were sourced from the GFMT short form consisting of controlled images taken on the same day (Burton et al., 2010), the EFCT dataset consisting of fairly high resolution images with variation in illumination and expression (White, Phillips, et al., 2015), the MFMT comprising of images of male models with variation in pose, expression, illumination and quality (Dowsett & Burton, 2015) and the CWT (Towler et al., 2019), a more challenging face matching test of uncontrolled images created by the training

agency for validation purposes. This provided a wide range of face imagery from different capture conditions and various levels of difficulty. The two face images in a pair were presented simultaneously as a single image with a resolution of 800 pixels wide and 560 pixels high, the face images were surrounded by white space. Representative images for the four different sources are shown in Figure 5.

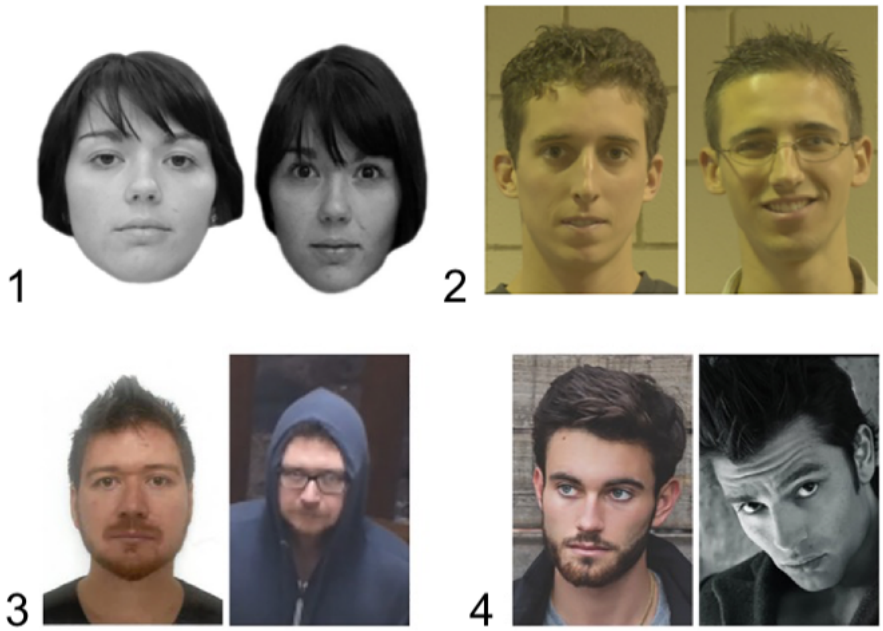


*Figure 5 - Example face matching stimuli used in the training evaluation, taken from Towler et al. (2019) (1 - GFMT, 2 - EFCT, 3 images representative of the CWT, 4 - MFMT)*

### 5.2.3. Procedure

The two-day training course delivered to police trainees closely resembled course D used by Towler et al. (2019) but was restricted in duration to two days due to constraints on the availability of the training venue. The course was delivered face-to-face by two trained face examiners who are subject-matter experts in the field. The training course covered seven different modules relevant to applied face matching, listed below (for a detailed list of topics used in the course see Appendix B).

1. An introduction to facial identification
2. Instruction in feature-based morphological face matching
3. Imaging and environmental factors affecting face matching
4. Applied uses of automated facial recognition technology
5. Human factors and bias in face matching
6. Process and procedures involved in applied face matching
7. Facial growth and development

The training also incorporated various face-matching tasks, including one-to-one matching and one-to-many matching, with feedback provided to trainees immediately after completing each task, in order to encourage them to apply the information learnt in the taught training modules and to help develop perceptual expertise in face matching through repeated practice (White et al., 2014). Trainees were also encouraged to work in pairs on some tasks in an attempt to further encourage training effects as demonstrated by Dowsett & Burton (2015).

Trainee participants completed trial A up to one week prior to training and trial B up to one week after completion of the training course. Control participants completed trial A and trial B on different days up to two weeks apart. Trials were delivered online using Qualtrics. Prior to completing the trials participants consented to take part in the study (see Appendix C) and provided some basic demographic information (age range and gender). Participants were then shown face images in side-by-side pairs in a randomised order and asked to respond whether the faces were a match or a non-match. Participants were also asked to provide a confidence rating for their decision on a four-point Likert scale ranging from '1 – Not confident' to '4 – extremely confident' (see Appendix D). All trials were self-paced with no time limit. Face matching response data was analysed using R (R Core Team, 2019). The study received a favourable opinion from the ethical committee of the Open University.

## 5.3. Results

### 5.3.1. Preliminary analysis

An initial analysis of trainee and control performance on trial A revealed that the trainee group had a median correct response score of 85 out of 107 (range = 71 to 101) whereas the control group had a median correct response score of 91 out of a possible 107 (range = 74 to 103). Based on raw scores the control group are slightly more proficient at face matching than the trainees prior to training but both groups have a considerable range in scores.

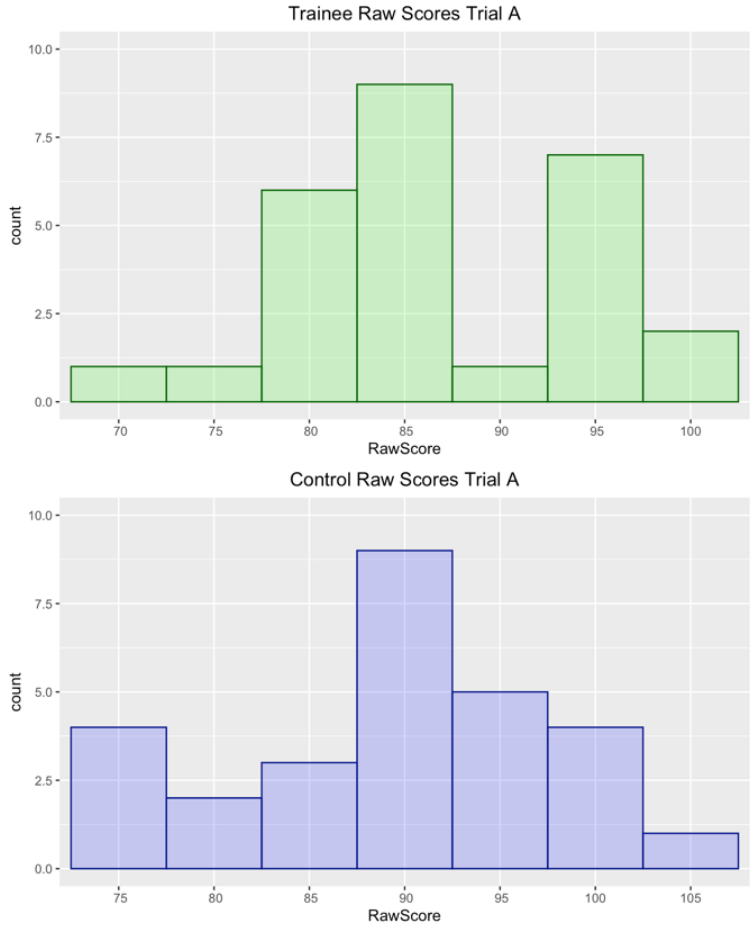


*Figure 6 - Distribution of raw scores on trial A for trainees and controls*

Due to some participants achieving maximum hit rates and correct rejections, as well as the raw scores for trainees appearing non-normally distributed (see Figure 6), subsequent signal detection analysis of the data was done using non-parametric measures of sensitivity (*A)* and bias (*b*) (Zhang & Mueller, 2005).

Figure 7 shows the correlation between performance on trial A and trial B for the control group. Performance is well correlated between the two trials. This indicates that there is a positive relationship between the two trials and that they are appropriate as a tool to demonstrate changes in face-matching accuracy and behaviour.

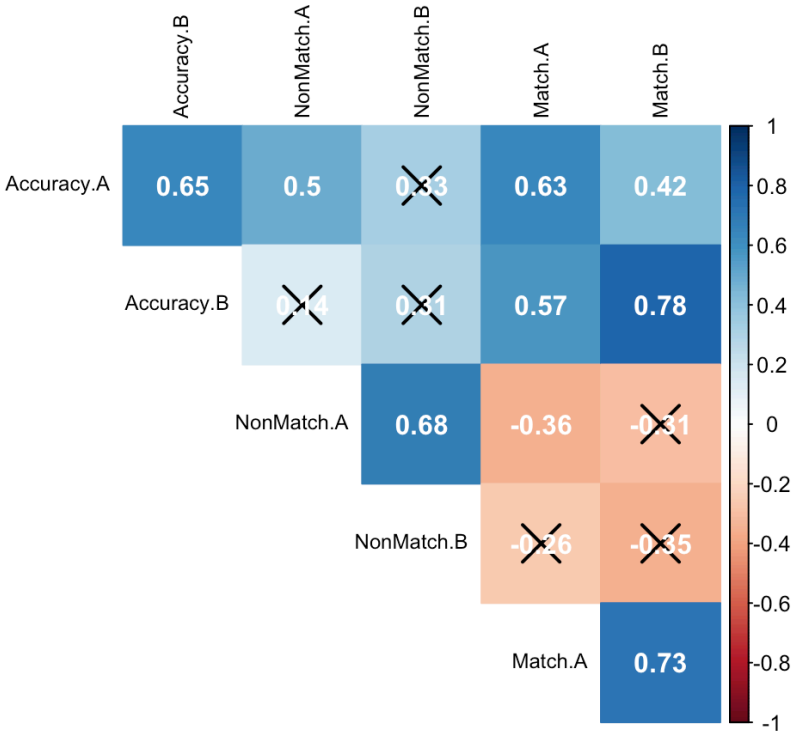


*Figure 7 - Correlation coefficients for overall accuracy, match and non-match scores between trial A and trial B for the control group (values with black cross are non-significant at the 95% confidence level)*

Match and non-match accuracy within and between trials show a weak negative correlation, replicating the dissociation between matching and non-matching performance for unfamiliar faces observed previously (Megreya & Burton, 2007). Therefore, as well as using signal

detection measures, participant performance on matching and non-matching trials were analysed separately alongside overall accuracy.

### 5.3.2. Overall accuracy

Summary statistics of overall accuracy for trainee and control groups on trial A and trial B are shown in Table 15. Although broadly comparable, the trainee group on average had slightly lower face matching accuracy prior to training than the control group. Trainee accuracy increased slightly on trial B, however performance post-training was still well within the range of control group accuracy.

*Table 15 - Summary statistics for trainee and control overall accuracy*

| Overall accuracy | | | | | | |
|---|---|---|---|---|---|---|
| Trainee | | | | | | |
| | Mean | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Trial A | 81.45 (7.38) | 79.44 | 66.36 | 76.64 | 88.79 | 94.39 |
| Trial B | 83.16 (6.93) | 84.26 | 69.44 | 79.17 | 87.04 | 95.37 |
| | | | | | | |
| Control | | | | | | |
| | Mean | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Trial A | 83.61 (7.47) | 85.05 | 69.16 | 79.21 | 88.79 | 96.26 |
| Trial B | 83.23 (8.33) | 85.65 | 68.52 | 76.16 | 89.81 | 94.44 |

Figure 8 shows the distribution of overall accuracy for trainee and control groups, with the black dots representing the scores of individual participants. Both groups displayed a wide range in individual scores on both trials indicating, that the training course had little if any impact on homogenising individual differences in face matching accuracy for the trainee group. Accuracy does not appear to be normally distributed for either group, which was confirmed with a Shapiro-Wilk test for normality ($W = 0.963$, $p = .004$), however variance

between the two groups appears to be equivalent, confirmed with Levene's test for homogeneity of variance ($F(1,53) = 0.19$, $p = .664$).

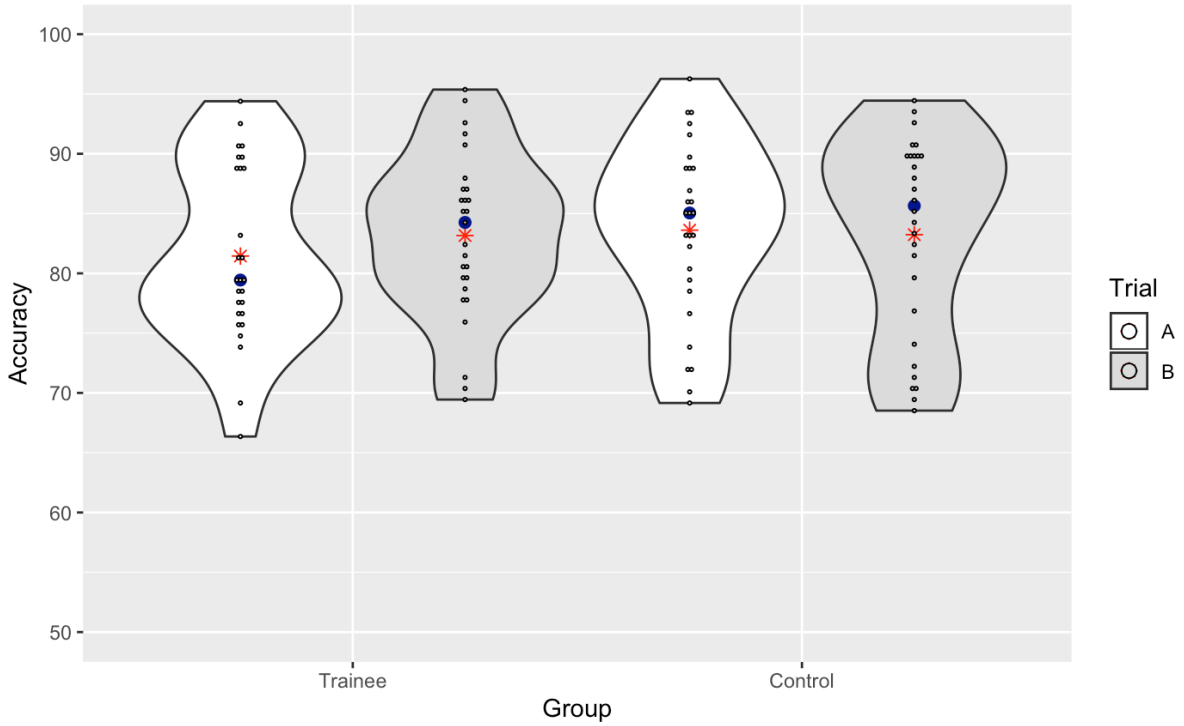


*Figure 8 - Trainee and control accuracy by trial (blue dot represents median group score and red star is the mean group score)*

Accuracy data was transformed by Tukey's ladders of powers using the rcompanion package (Mangiafico, 2019) prior to statistical tests[9]. A factorial ANOVA of the Tukey transformed data revealed no significant main effect for group ($F(1,53) = 0.43$, $\eta_p^2 = .007$, $p = .513$) or trial type ($F(1,53) = 0.93$, $\eta_p^2 = .002$, $p = .339$). Similarly, the interaction between group and trial type was non-significant ($F(1,53) = 1.70$, $\eta_p^2 = .004$, $p = .198$). These findings demonstrate that there was no significant difference in overall accuracy between the trainee

[9] A factorial ANOVA of the untransformed data was conducted for the purpose of comparison and yielded similarly non-significant results.

and control groups on either trial A or trial B. This finding is consistent with previous evidence that short training courses do not improve overall face matching accuracy.

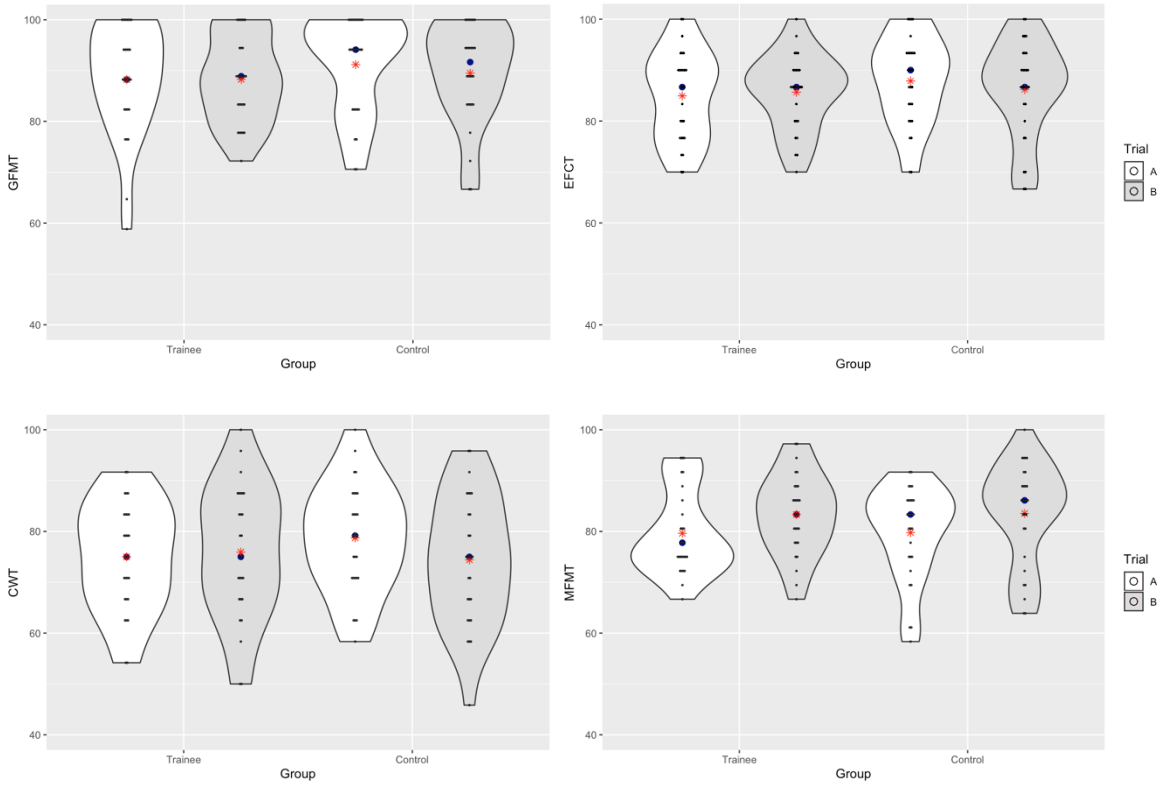


*Figure 9 - Trainee and control accuracy by trial on GFMT, EFCT, CWT and MFMT (blue dot represents median group score and red star is the mean group score)*

Figure 9 shows trainee and control accuracy for each of the four face matching tests used in trials A and B. As for overall accuracy, there is a wide range in individual accuracy across all tests and overlap in the distribution of scores between the trainee and control groups on both trial A and trial B. A significant main effect for trial was found on the MFMT ($F(1,53) = 14.23$, $\eta_p^2 = .004$, $p <. 001$) but not for group ($F(1,53) = 0.005$, $\eta_p^2 < .001$, $p =. 939$). Pairwise tests revealed that both trainees ($t(53) = -2.62$, $d = 0.391$, $p = .011$) and controls ($t(53) = -2.72$, $d = 0.459$, $p = .008$) were significantly more accurate at the group level on MFMT images in trial B compared to MFMT images in trial A. Because the effect was observed in both trainee and control groups it is not possible to attribute the improvement in accuracy

to training. There were no other significant interactions between group and/or trial for accuracy on any of the three remaining tests (GFMT, EFCT and CWT).

### 5.3.3. Match and non-match accuracy

Because of the dissociation in accuracy between matching and non-matching face pairs, these measures were also analysed separately to determine if training had any effect on matching or non-matching accuracy. Table 16 shows summary statistics for trainee and control match accuracy.

*Table 16 - Summary statistics for trainee and control match accuracy*

| Match accuracy | | | | | | |
|---|---|---|---|---|---|---|
| Trainee | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Trial A | 82.24 (10.59) | 83.33 | 59.26 | 75.00 | 88.89 | 96.30 |
| Trial B | 86.80 (7.39) | 87.27 | 70.91 | 83.64 | 93.64 | 96.36 |
| | | | | | | |
| Control | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Trial A | 84.72 (12.79) | 87.04 | 46.30 | 82.87 | 92.59 | 100.00 |
| Trial B | 84.87 (12.97) | 88.18 | 47.27 | 81.36 | 92.73 | 100.00 |

Trainee match accuracy increased on trial B compared to trial A and this increase was not observed in the control group. A closer inspection of the summary statistics suggests these increases occurred predominantly in the lower range of scores for the minimum and first quartile. This is confirmed in Figure 10, which shows the distribution of match accuracy for each individual participant. A visual inspection of Figure 10 reveals that the lower tail of the trainee match distribution is closer to the median value on trial B, however the upper tail remains largely the same.

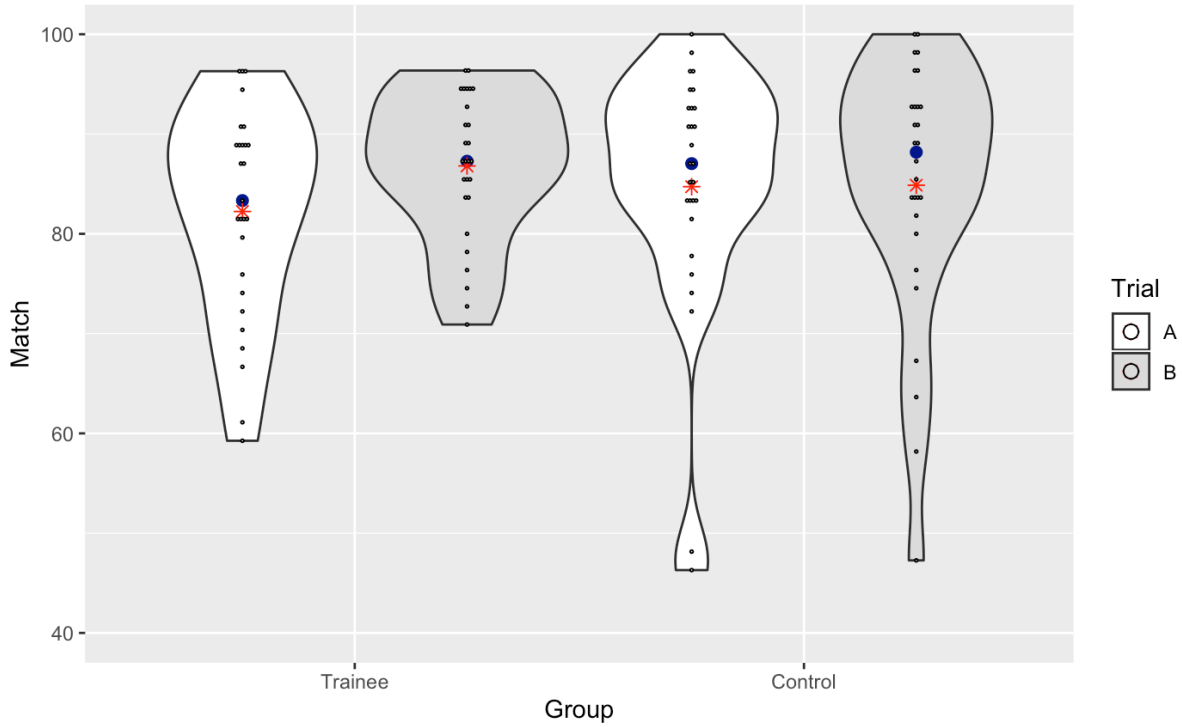


*Figure 10 - Trainee and control match accuracy by trial (blue dot represents median group score and red star is the mean group score)*

The match accuracy distributions are non-normally distributed ($W = 0.897$, $p < .001$) but do not show significant deviation in variance ($F(1,53) = 1.43$, $p = .237$). The data was Tukey transformed giving an approximately normal distribution ($W = 0.984$, $p = .214$), prior to a factorial ANOVA. No significant main effect was found for group ($F(1,53) = 0.42$, $\eta_p^2 = .006$, $p = .518$) or trial type ($F(1,53) = 2.91$, $\eta_p^2 = .012$, $p = .094$). Similarly, the interaction between group and trial type was non-significant ($F(1,53) = 1.87$, $\eta_p^2 = .008$, $p = .177$). Despite an increase in match accuracy at the lower end of the trainee distribution post training, this does not result in a significant improvement in performance for match pairs at the group level.

*Table 17 - Summary statistics for trainee and control non-match accuracy*

| Non-match accuracy | | | | | | |
|---|---|---|---|---|---|---|
| Trainee | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Trial A | 80.64 (10.02) | 83.02 | 56.60 | 71.70 | 88.68 | 100.00 |
| Trial B | 79.39 (14.45) | 84.91 | 47.17 | 70.75 | 89.62 | 100.00 |
| | | | | | | |
| Control | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Trial A | 82.48 (12.23) | 85.85 | 52.83 | 73.11 | 91.04 | 100.00 |
| Trial B | 81.54 (13.19) | 82.08 | 45.28 | 77.36 | 91.04 | 100.00 |

Summary statistics for non-match accuracy (Table 17) show an increase in the range of non-match accuracy for trainees post training, which is greater than that of the control group. However, there does not appear to be a notable difference in overall non-match accuracy for either group, this can be observed in the non-match distributions in Figure 11. A factorial ANOVA of the Tukey transformed data revealed no main effect for non-match accuracy by group ($F(1,53) = 0.53$, $\eta_p^2 = .008$, $p = .471$), trial type ($F(1,53) = 0.12$, $\eta_p^2 < .001$, $p = .728$) or the interaction between group and trial type ($F(1,53) = 0.09$, $\eta_p^2 < .001$, $p = .770$).

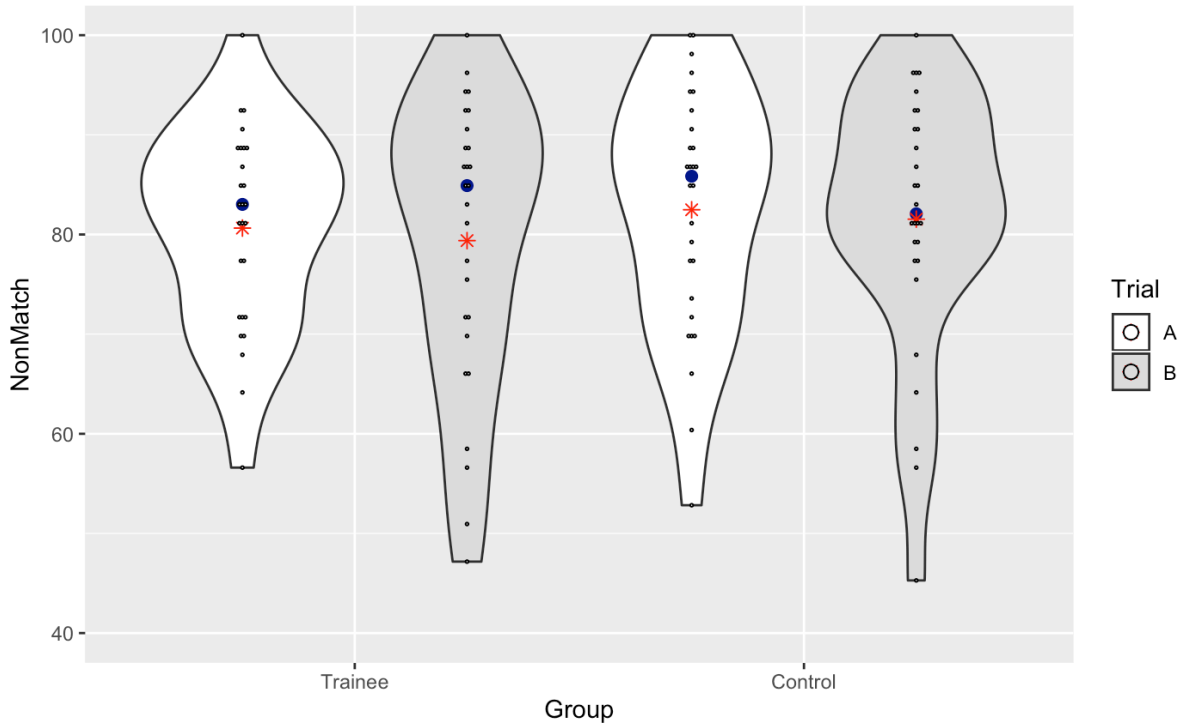


*Figure 11 - Trainee and control non-match accuracy by trial (blue dot represents median group score and red star is the mean group score)*

### 5.3.4. Sensitivity and bias

Signal detection analysis was conducted on the data to understand whether training affected sensitivity, (the ability to detect matches and reject non-matches) and response bias (the measure of how likely an individual is to respond match or non-match). Signal detection analysis generates a score for sensitivity based on the number of hits (correct matches) and false alarms (incorrect non-matches) that is free from bias (Bate et al., 2018). A separate measure is generated to indicate response bias. Sensitivity is often reported as d-prime (*d'*) alongside criterion (*c*) as a measure of bias. Both *d'* and *c* are calculated from the converted z-scores of hits and false alarms and assumes a normal distribution. Due to the non-normal distribution of overall accuracy, non-parametric measures of sensitivity (*A)* and (*b*) were used (Zhang & Mueller, 2005). *A* measures sensitivity within a range of 0 to 1. Bias or *b* indicates response bias, with a value of zero being a neutral response. A negative

value of *b* indicates that the individual is more likely to respond match and has a liberal response bias. Positive values of *b* indicate a greater propensity to respond non-match, which is a conservative response bias. Summary statistics for *A* and *b* are given in Table 18.

*Table 18 - Summary statistics for trainee and control sensitivity (A) and bias (b)*

| Trainee | | | | | | |
|---|---|---|---|---|---|---|
| Trial A | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Sensitivity *(A)* | 0.873 (0.062) | 0.863 | 0.722 | 0.831 | 0.935 | 0.972 |
| Bias (b) | -0.040 (0.319) | -0.006 | -0.472 | -0.346 | 0.143 | 0.593 |
| Trial B | | | | | | |
| | Mean | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Sensitivity *(A)* | 0.891(0.052) | 0.904 | 0.754 | 0.859 | 0.922 | 0.975 |
| Bias (b) | -0.148 (0.401) | -0.123 | -0.855 | -0.317 | 0.072 | 0.772 |
| Control | | | | | | |
| Trial A | | | | | | |
| | Mean | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Sensitivity *(A)* | 0.896 (0.056) | 0.906 | 0.751 | 0.865 | 0.935 | 0.982 |
| Bias (b) | -0.056 (0.449) | -0.146 | -0.745 | -0.323 | 0.186 | 1.124 |
| Trial B | | | | | | |
| | Mean | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Sensitivity *(A)* | 0.891(0.066) | 0.912 | 0.756 | 0.847 | 0.946 | 0.969 |
| Bias (b) | -0.082 (0.445) | -0.178 | -0.799 | -0.399 | 0.246 | 0.911 |

The trainee group had an increase in sensitivity (*A*) after training along with a decrease in bias (*b*) suggesting a slightly more liberal response bias at the group level. The full range of *b* values for the training group also increases on trial B suggesting a greater diversity in response bias after training.

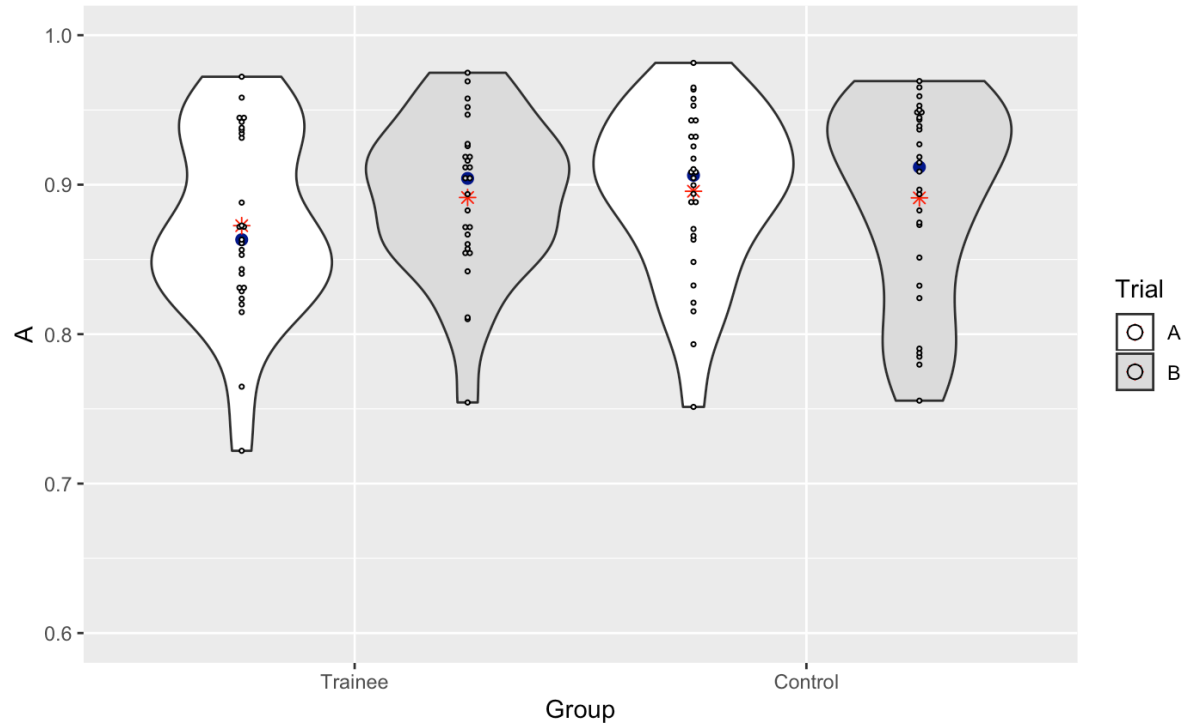


Figure 12 - Trainee and control sensitivity (A) by trial (blue dot represents median group score and red star is the mean group score)

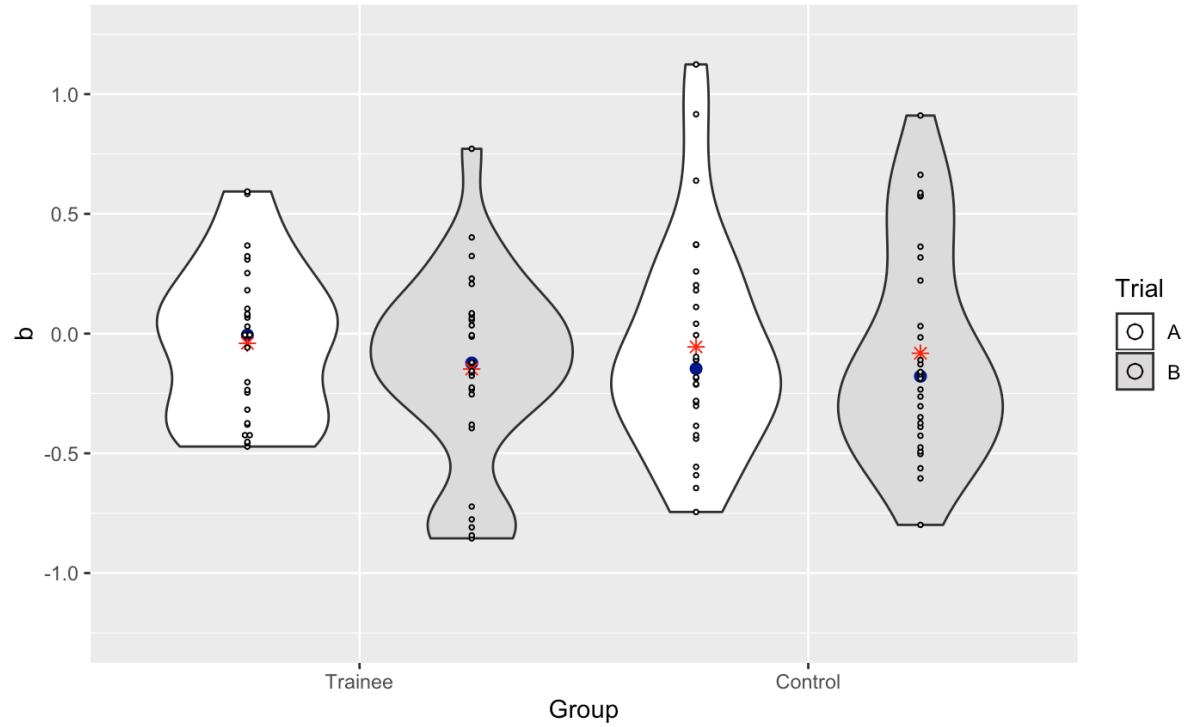


*Figure 13 - Trainee and control bias (b) by trial (blue dot represents median group score and red star is the mean group score)*

The distributions of *A* and *b* values for trainees and controls are shown in Figure 12 and Figure 13 respectively. Despite a small increase in trainee sensitivity post training there is still a large overlap in values of *A* for trainees and controls across both trials. Regarding *b,* the distribution of values for controls are visually similar. For the trainee group there appears to be five individuals who have a much more liberal response bias on trial B than for trial A, who are likely causing the decrease in average response bias for the trainee group post training. Interestingly there is one individual who has a much more conservative response bias after training, resulting in wider range of values of *b* for trainees on trial B.

A factorial ANOVA of *A* following Tukey transformation revealed no main effect for group ($F(1,53) = 0.76$, $\eta_p^2 = .012$, $p = .387$) or trial ($F(1,53) = 1.47$, $\eta_p^2 = .004$, $p = .231$) and no interaction between trial and group ($F(1,53) = 2.59$, $\eta_p^2 = .007$, $p = .114$). Similarly, for *b* there was no main effect for group ($F(1,53) = 0.07$, $\eta_p^2 = .001$, $p = .799$) or trial ($F(1,53) = 2.06$, $\eta_p^2 = .007$, $p = .157$) and no interaction between group and trial ($F(1,53) = 0.75$, $\eta_p^2 = .003$, $p = .389$). Despite changes in the response bias and sensitivity of some trainees post training these differences were not significant at the group level.

Figure 14 plots the relationship between *b* and *A* for trainees and controls. On trial A there is no relationship between sensitivity and response bias for trainees, $r(27) = -.07$, $p = .716$, or controls, $r(28) = .05$, $p = .791$. For trial B there was similarly no relationship observed for the control group, $r(28) = .00$, $p = .994$. However, for the trainee group a positive relationship with moderate effect size was present between *b* and *A* on trial B, $r(27) = 0.48$, $p = .012$. This indicates that after training individuals with low sensitivity are more likely to respond match than prior to training.

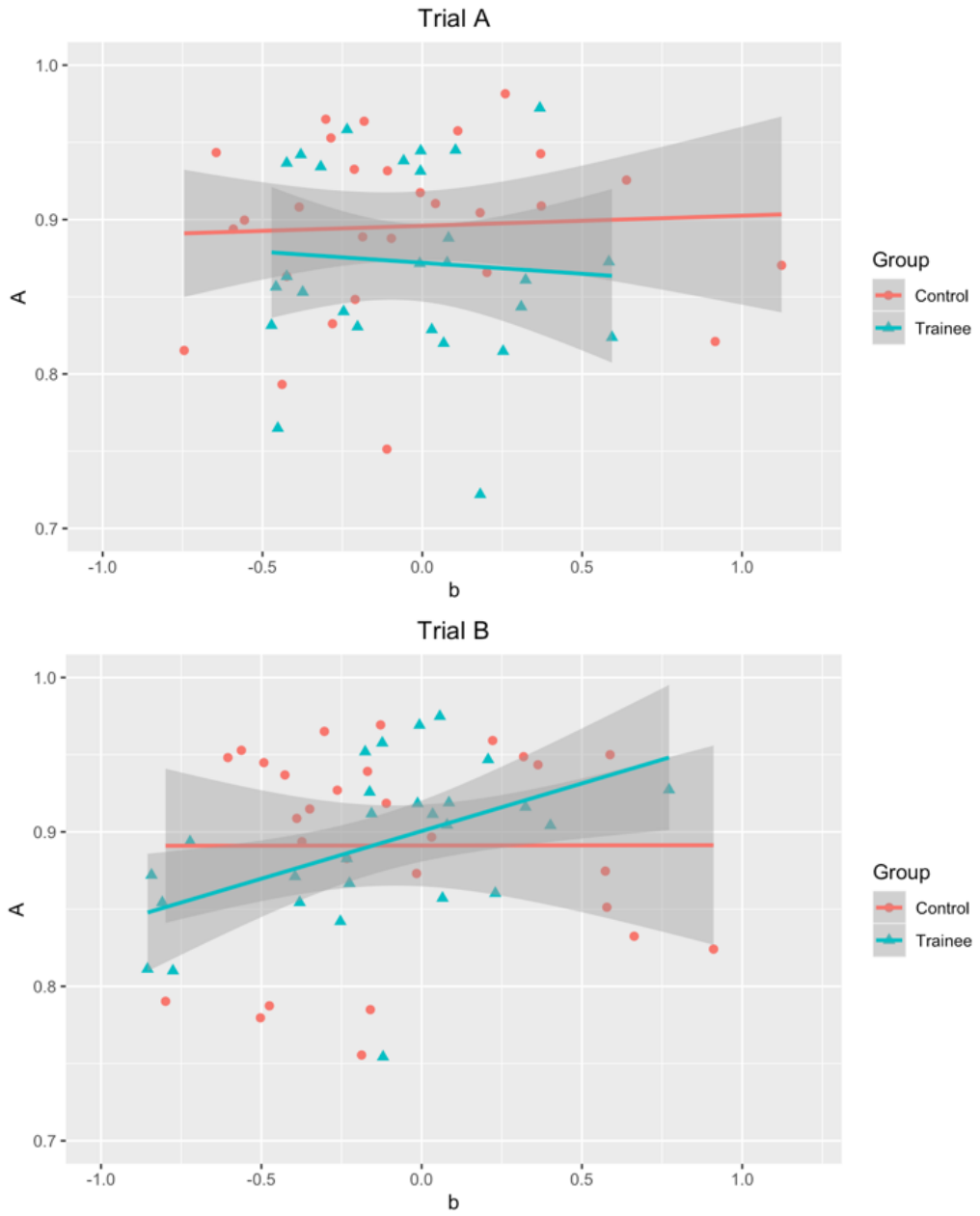


*Figure 14 - Relationship between bias (b) and sensitivity (A) for trainees and controls on trial A and trial B*

### 5.3.5. High performers and low performers

To further investigate the change in sensitivity and bias for low performing trainees the dataset was split into high and low performers. The ten individuals with the highest sensitivity scores and ten individuals with the lowest sensitivity scores on trial A were taken

from the trainee and control groups to form four new groups. Summary statistics of *A* and *b* are shown for trainees in Table 19 - Summary statistics for low trainee and high trainee sensitivity (A) and bias (b) and controls in Table 20.

*Table 19 - Summary statistics for low trainee and high trainee sensitivity (A) and bias (b)*

| Trial A | | | | | | |
|---|---|---|---|---|---|---|
| Low trainee | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Sensitivity *(A)* | 0.812 (0.038) | 0.826 | 0.722 | 0.816 | 0.831 | 0.836 |
| Bias (b) | 0.006 (0.346) | 0.048 | -0.472 | -0.235 | 0.235 | 0.593 |
| | | | | | | |
| High trainee | | | | | | |
| | Mean | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Sensitivity *(A)* | 0.939 (0.022) | 0.940 | 0.888 | 0.934 | 0.945 | 0.972 |
| Bias (b) | -0.087 (0.249) | -0.032 | -0.424 | -0.297 | 0.060 | 0.368 |
| | | | | | | |
| Trial B | | | | | | |
| Low trainee | | | | | | |
| | Mean | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Sensitivity *(A)* | 0.856 (0.029) | 0.859 | 0.810 | 0.845 | 0.870 | 0.905 |
| Bias (b) | -0.317 (0.390) | -0.243 | -0.855 | -0.681 | -0.007 | 0.229 |
| | | | | | | |
| High trainee | | | | | | |
| | Mean | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Sensitivity *(A)* | 0.936 (0.025) | 0.927 | 0.904 | 0.916 | 0.956 | 0.975 |
| Bias (b) | 0.087 (0.292) | 0.014 | -0.176 | -0.095 | 0.078 | 0.772 |

The average score of *A* for low trainees show a slight increase on trial B along with a similarly modest decrease in range. This is accompanied by a shift from an almost neutral average score for *b* on trial A to a negative average value for *b* on trial B. For high performing trainees there is a very small decrease in *A* scores on trial B with *b* scores showing a slight shift towards conservatism on trial B. For controls, low performers show a slight increase in

*A* on trial B, whereas high performers decrease slightly. Average scores for *b* are very similar between trial A and B for low performing controls with a slight shift towards negative values and high performing controls show the opposite and shift very slightly towards positive values for *b*.

*Table 20 - Summary statistics for low control and high control sensitivity (A) and bias (b)*

| Trial A | | | | | | |
|---|---|---|---|---|---|---|
| Low control | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Sensitivity *(A)* | 0.835 | 0.840 | 0.751 | 0.817 | 0.865 | 0.888 |
| Bias (b) | -0.007 | -0.161 | -0.745 | -0.388 | 0.127 | 1.125 |
| High control | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Sensitivity *(A)* | 0.949 | 0.948 | 0.926 | 0.935 | 0.962 | 0.982 |
| Bias (b) | -0.036 | -0.146 | -0.645 | -0.268 | 0.222 | 0.639 |
| Trial B | | | | | | |
| Low control | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Sensitivity *(A)* | 0.839 | 0.828 | 0.756 | 0.788 | 0.889 | 0.953 |
| Bias (b) | -0.072 | -0.280 | -0.799 | -0.496 | 0.438 | 0.911 |
| High control | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Sensitivity *(A)* | 0.935 | 0.947 | 0.851 | 0.940 | 0.957 | 0.969 |
| Bias (b) | 0.074 | 0.047 | -0.492 | -0.217 | 0.352 | 0.588 |

Figure 15 and Figure 16 show the distribution of *A* and *b* scores respectively, split by group and trial.

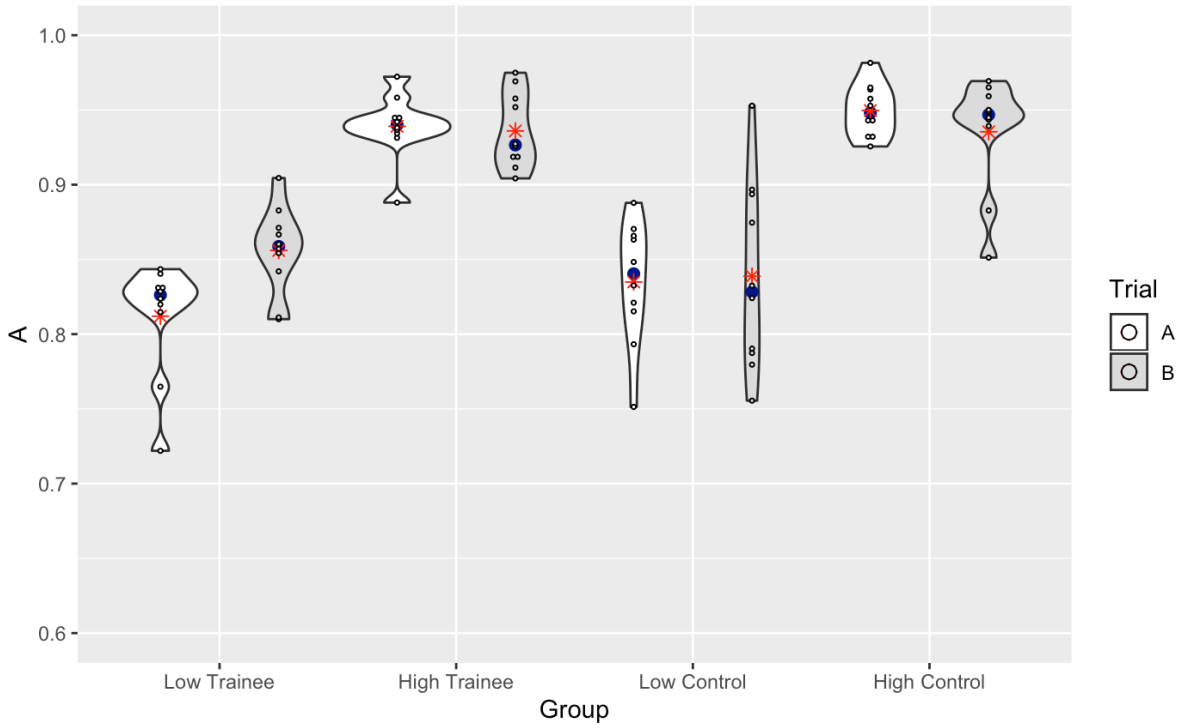


*Figure 15 - High and low performer A scores by group and trial (blue dot represents median group score and red star is the mean group score)*

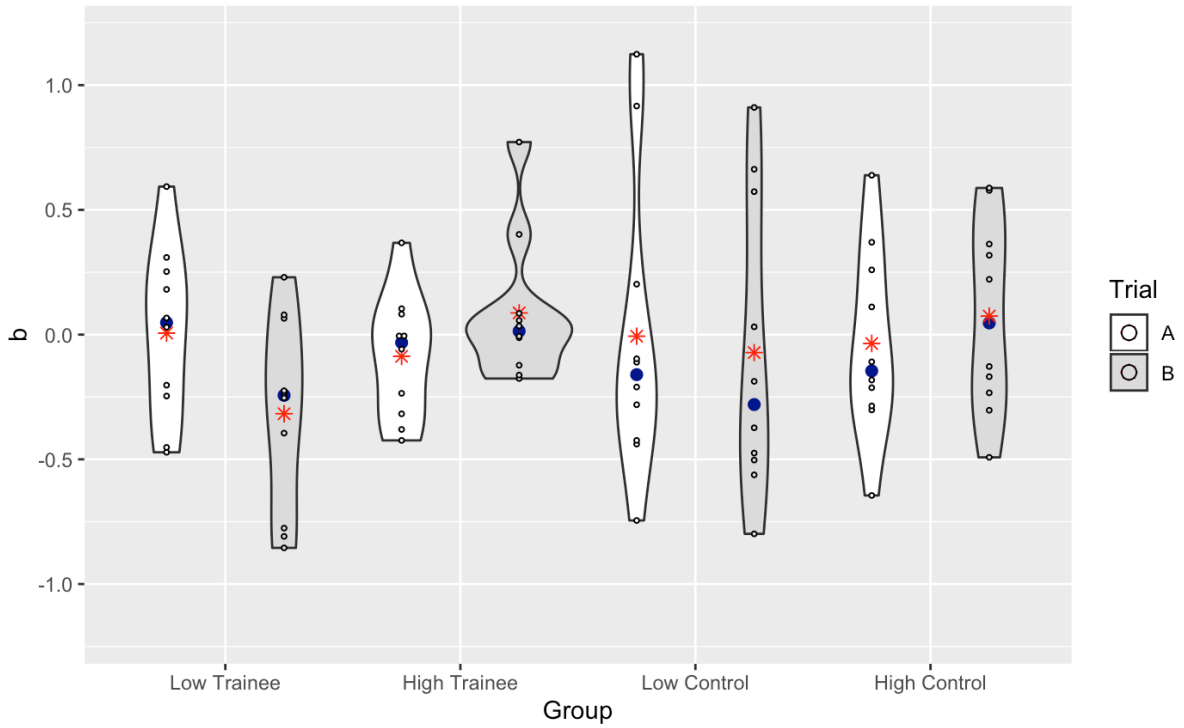


*Figure 16 - High and low performer b scorers by group and trial (blue dot represents median group score and red star is the mean group score)*

All data were Tukey transformed due to non-normal distributions prior to a series of ANOVAs. For *A* there was a significant effect for group ($F(3,36) = 47.55$, $\eta_p^2 = .717$, $p < .001$) but not for trial ($F(1,36) = .84$, $\eta_p^2 = .008$, $p = .367$). Pairwise tests for group, as expected, found no reliable difference between high performing trainees and controls or low performing trainees and controls on both trials (all tests $p = 1$). All differences between high performers and low performers within and between groups were significant (all tests $p < .001$). There was no reliable interaction between trial and group ($F(3,36) = 2.81$, $\eta_p^2 = .078$, $p = .053$). These findings indicate that for sensitivity the difference between high and low performers remained largely consistent for both groups across trial A and trial B. Although the low performing trainee group did show an increase in *A* on trial B this was not a reliable improvement and the group still had significantly lower sensitivity than high performers post training on trial B (see Figure 15).

For *b* there was no main effect for group ($F(3,36) = 0.39$, $\eta_p^2 = .028$, $p = .760$) or trial ($F(1,36) = 0.33$, $\eta_p^2 = .001$, $p = .569$) but a reliable interaction was found between group and trial ($F(3,36) = 5.97$, $\eta_p^2 = .055$, $p = .002$). Post hoc pairwise tests revealed that this interaction was significant with a moderate effect size for the low performing trainee group ($t(36) = 3.55$, $d = 0.391$, $p = .001$) but not for any of the other three groups. Post-training the low trainees had a significantly more liberal response bias (i.e. tendency to respond match) indicating a change in face matching behaviour not observed in other groups (see Figure 16). However, this change in response bias resulted in only a modest increase in sensitivity for the low performing trainees, which was not found to be a reliable improvement.

Match and non-match accuracy for high and low performers were analysed and compared. Summary statistics for high and low performing trainees are shown in Table 21. Low performing trainees had a marked increase in match accuracy post training and a slight decrease in non-match accuracy, which was most pronounced for the lower quartile. For

high performing trainees match accuracy decreases very slightly post training. For non-match accuracy there is a slight increase in accuracy, with the top-performing trainee at ceiling in both trial A and B, however it may not be possible to detect any effects of training for this group due to their initial high performance.

*Table 21 - Summary statistics for low performing and high performing trainee match accuracy and non-match accuracy*

| Trial A | | | | | | |
|---|---|---|---|---|---|---|
| Low trainee | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Match | 74.26 (9.42) | 75.00 | 59.26 | 68.98 | 81.48 | 87.04 |
| Non match | 74.53 (9.55) | 74.53 | 56.60 | 71.70 | 80.19 | 88.68 |
| | | | | | | |
| High trainee | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Match | 91.30 (4.71) | 90.74 | 81.48 | 88.89 | 95.83 | 96.30 |
| Non match | 88.11 (5.56) | 88.68 | 81.13 | 83.49 | 90.09 | 100.00 |
| | | | | | | |
| Trial B | | | | | | |
| Low trainee | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Match | 86.00 (6.30) | 86.36 | 74.55 | 83.64 | 90.45 | 94.55 |
| Non match | 70.38 (14.06) | 73.58 | 47.17 | 60.38 | 80.19 | 86.79 |
| | | | | | | |
| High trainee | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Match | 87.45 (8.29) | 89.09 | 70.91 | 85.45 | 94.55 | 94.55 |
| Non match | 90.94 (4.77) | 89.62 | 84.91 | 87.26 | 93.87 | 100.00 |

For controls (Table 22), match accuracy of low performers is roughly equivalent on trial A and trial B, with the exception of one individual who is at ceiling on trial B (Figure 17). For non-match accuracy there is slight overall decrease in scores, however this group shows the largest range in performance for all four groups indicating large individual differences in

how low performing controls respond to matches and non matches. High performing controls decreased slightly in match accuracy on trial B but were roughly equivalent for non-match accuracy on both trials.

*Table 22 - Summary statistics for low performing and high performing control match accuracy and non-match accuracy*

| Trial A | | | | | | |
|---|---|---|---|---|---|---|
| Low control | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Match | 75.74 (15.99) | 83.33 | 46.30 | 73.15 | 84.72 | 90.74 |
| Non match | 75.28 (14.82) | 71.70 | 52.83 | 66.98 | 83.96 | 100.00 |
| | | | | | | |
| High control | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Match | 91.48 (7.72) | 92.59 | 74.07 | 90.74 | 96.30 | 100.00 |
| Non match | 90.38 (6.92) | 89.62 | 77.36 | 86.79 | 95.75 | 100.00 |
| | | | | | | |
| Trial B | | | | | | |
| Low control | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Match | 78.00 (16.25) | 83.64 | 47.27 | 69.09 | 89.09 | 100.00 |
| Non match | 74.34 (17.21) | 78.30 | 45.28 | 59.91 | 89.15 | 94.34 |
| | | | | | | |
| High control | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Match | 87.45 (10.44) | 88.18 | 63.64 | 84.09 | 95.45 | 98.18 |
| Non match | 90.57 (7.23) | 91.51 | 77.36 | 87.26 | 96.23 | 100.00 |

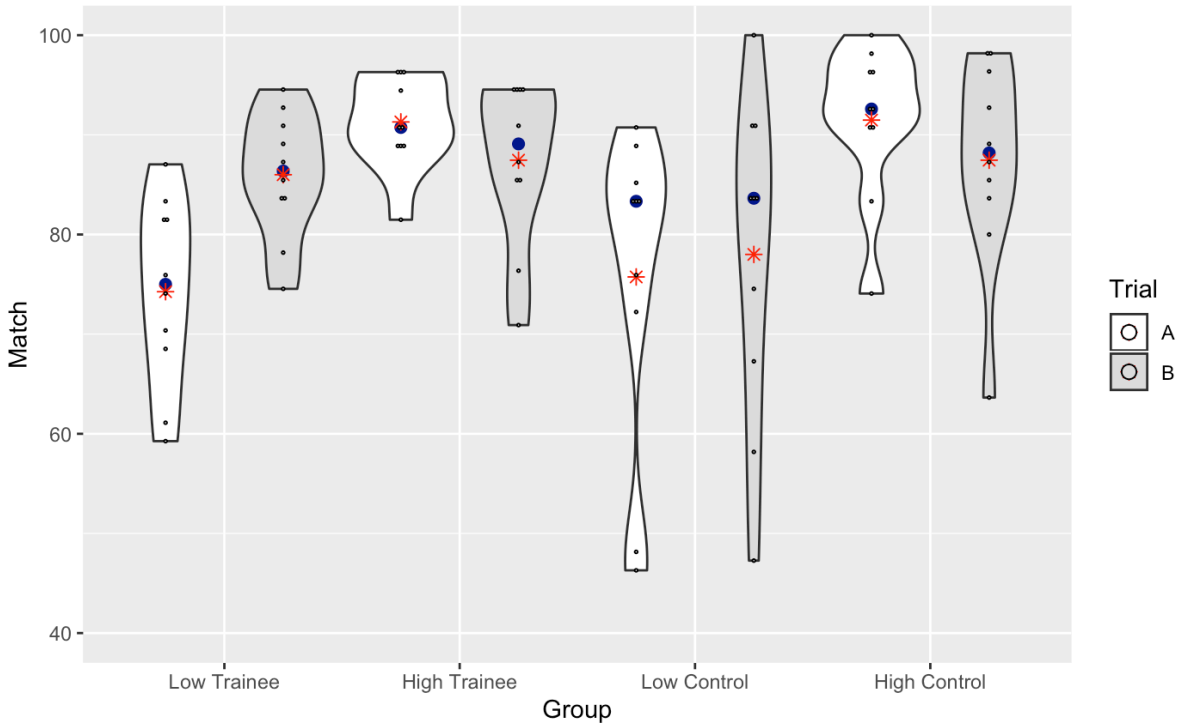


*Figure 17 - High and low performer match scores by group and trial (blue dot represents median group score and red star is the mean group score)*

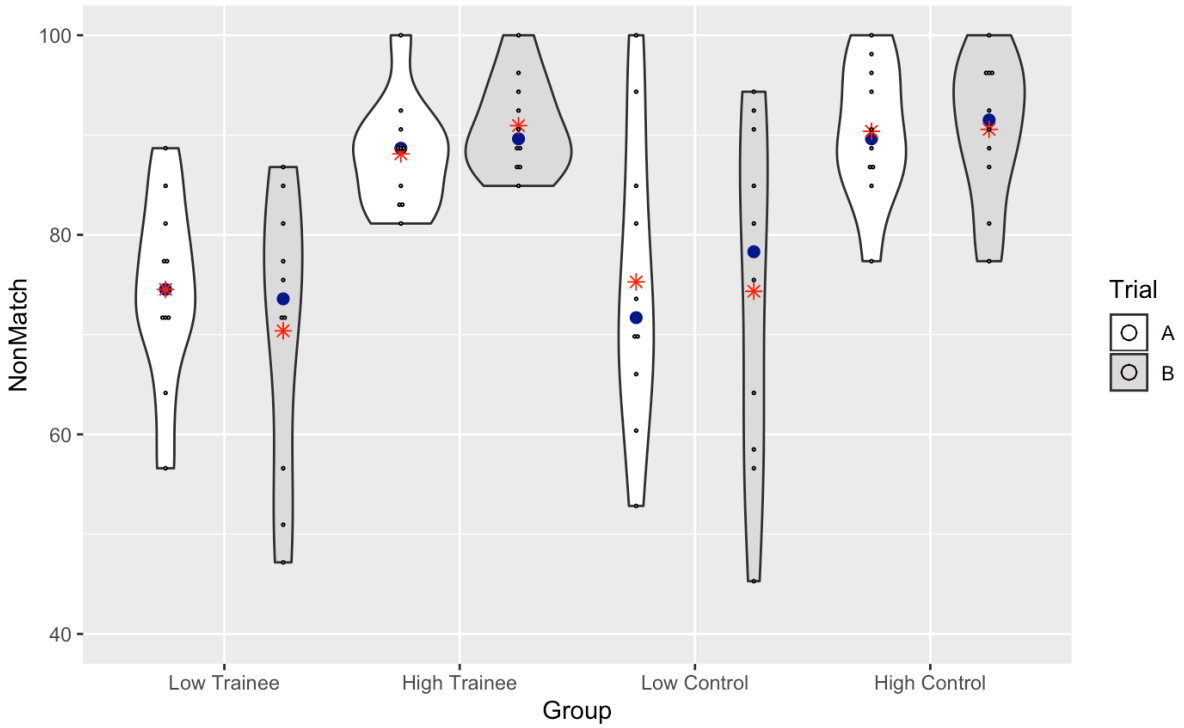


*Figure 18 - High and low performer non-match scores by group and trial (blue dot represents median group score and red star is the mean group score)*

Figure 17 and Figure 18 respectively show the distributions of match and non-match performance across the four groups for trial A and B. The low performing trainees show the largest change in match accuracy distribution and were at a comparable level to high performers on trial B but not on trial A. The high performers show a slight decrease in match accuracy, but this could be explained as regression to the mean. For non-match accuracy both low performing groups show a decrease in the maximum and minimum scores but similar average scores.

All data were Tukey transformed due to non-normal distributions prior to a series of factorial ANOVAs. For match accuracy there was a significant effect for group ($F(3,36) = 4.94$, $\eta_p^2 = .260$, $p = .006$) but not for trial ($F(1,36) = .79$, $\eta_p^2 = .003$, $p = .380$). A significant interaction was found between trial and group ($F(3,36) = 9.71$, $\eta_p^2 = .105$, $p < .001$). Pairwise tests for groups were performed. On trial A significant differences were observed between high performers and low performers within and between the trainee and control groups. However, on trial B no significant differences were found between high performing trainees and low performing trainees or between high performing controls and low performing trainees (p values = 1), indicating that after training the low performing trainee group had match accuracy scores comparable to both high performing groups. Post hoc pairwise tests of trial type revealed a reliable difference for low performing trainee match accuracy with a large effect size ($t(36) = -4.78$, $d = 1.47$, $p < .001$). The difference in match accuracy between trial A and trial B for high performing controls was bordering on significance ($t(36) = 2.04$, $d = 0.44$, $p = .049$), however examination of the distribution of match scores in Figure 17 shows there is an individual outlier that may be driving the lower performance of the group on trial B.

For non-match accuracy there was a reliable effect for group ($F(3,36) = 10.19$, $\eta_p^2 = .260$, $p < .001$) but not for trial ($F(1,36) = 0.08$, $\eta_p^2 < .001$, $p = .776$) and no interaction between

group and trial ($F(3,36) = 0.64$, $\eta_p^2 = .011$, $p = .594$). Pairwise tests for group revealed the expected pattern of no significant differences between high performing trainees and controls or between low performing trainees and controls on both trials (p values = 1).

### 5.3.6. Confidence decisions

As well as responding match or non-match to a facial-image pair, respondents also had to rate their confidence in the decision on a four point Likert scale, ranging from 'Not confident' to 'Extremely confident'. The impact of training on confidence decisions was analysed. The distribution of confidence decisions for match and non-match pairs are shown for trainees in Figure 19 and controls in Figure 20. Overall trends in the distribution of confidence decisions appear similar between the groups. 'Not confident' decisions are the least prevalent, followed by 'Extremely confident' decisions. 'Quite confident' decisions are the most frequently used for both groups on both trials, indicating that individuals were less likely to use the extremes of the confidence scale. Due to differences in the number of match and non-match responses between the two groups on each trial it is difficult to interpret from these plots if there are any changes in confidence decisions on trial B, however the overall pattern of confidence decisions appears similar.

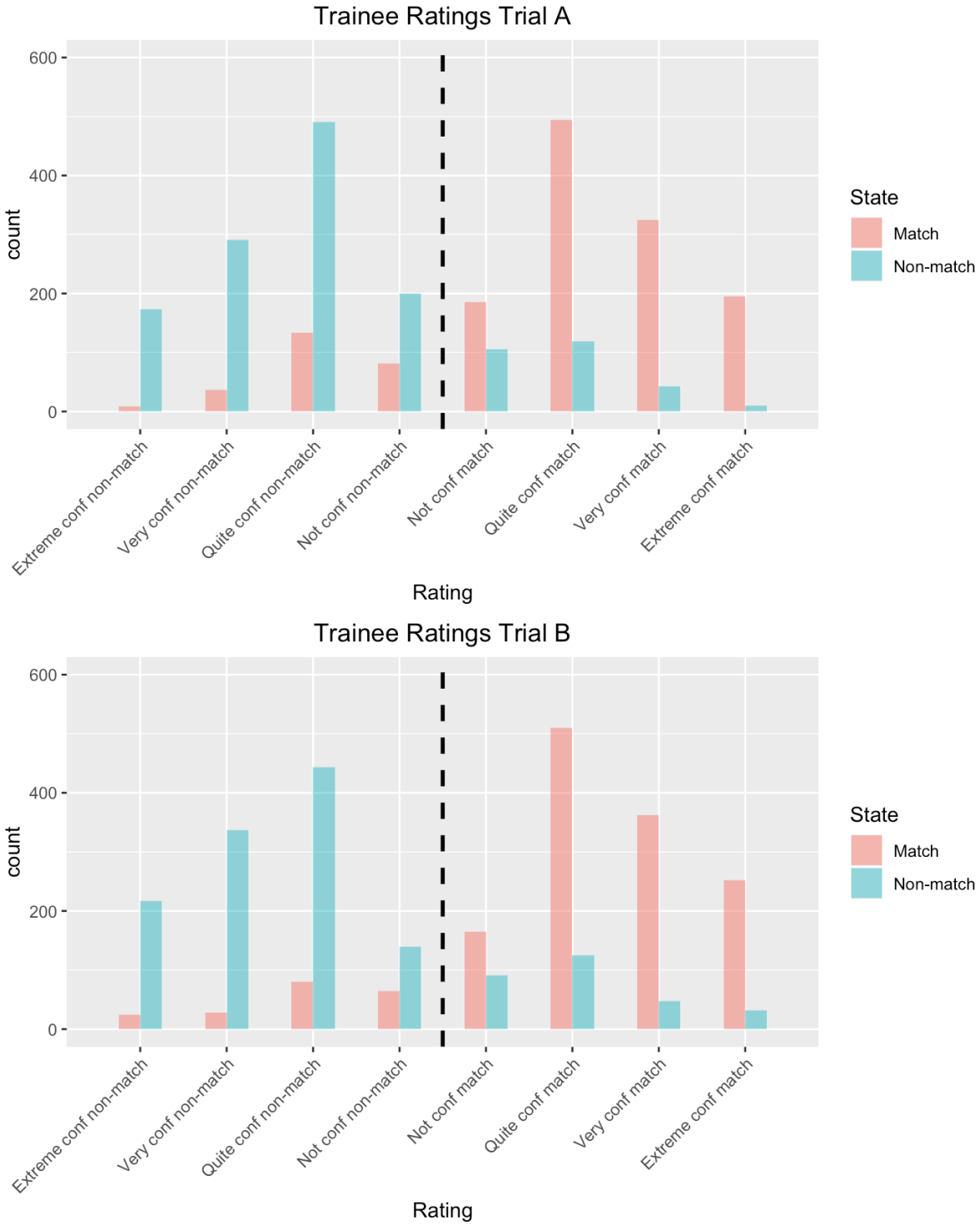


*Figure 19 - Distributions of trainee confidence decisions for trial A and trial B*

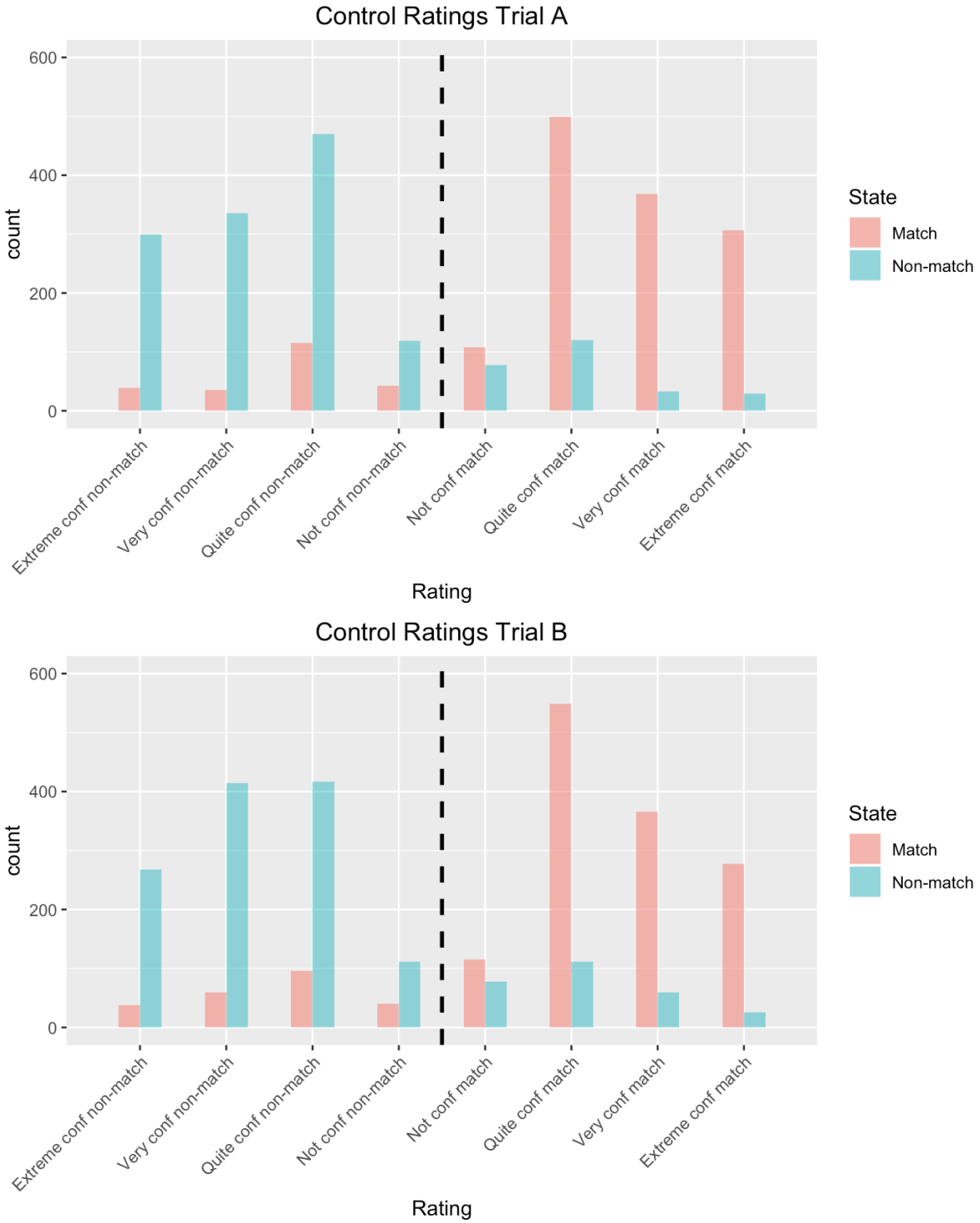


*Figure 20 - Distributions of control confidence decisions for trial A and trial B*

To better understand changes in confidence on trial B the differences in confidence decisions between trial B and trial A were calculated for each group. A positive median value reflects an overall increase in confidence rating use on trial B and a negative value shows a decrease in use. Results were separated between match decisions and non-match decisions. Summary statistics for the trainee group are shown in Table 23.

*Table 23 – Summary statistics of confidence decisions differences between trial B and trial A for trainee group*

| Trainee | | | | | |
|---|---|---|---|---|---|
| Match decisions | | | | | |
| | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Extremely confident | 0.0 | -10.0 | -0.5 | 4.5 | 48.0 |
| Very confident | 4.0 | -36.0 | -2.5 | 6.5 | 22.0 |
| Quite confident | 0.0 | -14.0 | -4.5 | 4.0 | 20.0 |
| Not confident | -3.0 | -12.0 | -5.0 | 4.0 | 9.0 |
| | | | | | |
| Non-match decisions | | | | | |
| | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Extremely confident | 0.0 | -11.0 | -1.0 | 3.5 | 43.0 |
| Very confident | 1.0 | -16.0 | -3.0 | 8.0 | 18.0 |
| Quite confident | -4.0 | -28.0 | -8.0 | 2.5 | 12.0 |
| Not confident | -4.0 | -17.0 | -7.0 | 1.5 | 9.0 |

The very high maximum value for ‘Extremely confident’ decision use, compared to the median value of zero suggests an outlier may be skewing the results. Closer inspection of the data revealed that one trainee shifted to predominantly making ‘Extremely confident’ decisions on trial B after training , and this shift was not observed for any other individual. Table 24 shows the summary statistics for trainees with the outlier removed, demonstrating that the increase in ‘Extremely confident’ decisions was isolated to this individual trainee.

*Table 24 – Summary statistics of confidence decisions differences between trial B and trial A for trainee group with outlier removed*

| Trainee (outlier removed) | | | | | |
|---|---|---|---|---|---|
| Match decisions | | | | | |
| | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Extremely confident | 0.0 | -10.0 | -0.8 | 3.5 | 14.0 |
| Very confident | 4.5 | -15.0 | -2.0 | 6.8 | 22.0 |
| Quite confident | 0.0 | -12.0 | -3.8 | 4.0 | 20.0 |
| Not confident | -2.5 | -12.0 | -5.0 | 4.5 | 9.0 |
| | | | | | |
| Non-match decisions | | | | | |
| | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Extremely confident | 0.0 | -11.0 | -1.5 | 2.5 | 15.0 |
| Very confident | 1.5 | -12.0 | -3.0 | 8.5 | 18.0 |
| Quite confident | -4.0 | -28.0 | -7.8 | 2.8 | 12.0 |
| Not confident | -4.0 | -17.0 | -7.0 | 1.8 | 9.0 |

There are large individual differences in the extent to which trainee confidence changed between trial A and trail B, reflected by the wide range in minimum and maximum values for each confidence decision. An outlier was also found in the control group, this individual had a substantial increase in very confident match decisions on trial B that was not observed for any other participant. Summary statistics of the changes in confidence for the control group are show in Table 25, with the outlier removed. As for the trainee group there are large individual differences in confidence between trial A and trial B for control participants.

*Table 25 – Summary statistics of confidence decisions differences between trial B and trial A for control group with outlier removed*

| Control (outlier removed) | | | | | |
|---|---|---|---|---|---|
| Match decisions | | | | | |
| | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Extremely confident | -1.0 | -16.0 | -3.5 | 1.0 | 24.0 |
| Very confident | -1.0 | -27.0 | -5.0 | 2.0 | 20.0 |
| Quite confident | 1.0 | -21.0 | -4.0 | 7.0 | 23.0 |
| Not confident | 0.0 | -17.0 | -1.5 | 2.5 | 14.0 |
| | | | | | |
| Non-match decisions | | | | | |
| | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Extremely confident | -1.5 | -14.0 | -5.3 | 2.0 | 21.0 |
| Very confident | 2.5 | -8.0 | -3.0 | 6.3 | 29.0 |
| Quite confident | -1.0 | -39.0 | -5.0 | 1.5 | 21.0 |
| Not confident | 0.0 | -11.0 | -1.3 | 1.0 | 12.0 |

Based on the distributions of confidence decisions and changes in frequency of confidence decisions, there does not appear to be a consistent change in confidence for match or non-match decisions after training, implying that training did not change confidence in face matching decisions.

Finally, the calibration of confidence decisions to facial image pair difficulty was analysed. The aim of this analysis was to understand if training resulted in confidence decisions that were more sensitive to the difficultly of a face-matching task. Average accuracy of all responses for each facial image pair in trial A and trial B were calculated, with trainee and controls scores collapsed into a single group. This resulted in a score of item difficulty for all facial image pairs. Spearman's rho ($r_s$) was used to calculate the correlation between the confidence decisions of individual participants and item difficulty for trial A and trial B.

*Table 26 - Summary statistics of Spearman's rho values of confidence-accuracy relationship for match pairs*

| Correlation coefficients for match pairs (Spearman's Rho) | | | | | | |
|---|---|---|---|---|---|---|
| Trainee | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Trial A | 0.318 (0.154) | 0.329 | 0.018 | 0.223 | 0.435 | 0.585 |
| Trial B | 0.439 (0.113) | 0.443 | 0.223 | 0.366 | 0.501 | 0.730 |
| | | | | | | |
| Control | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Trial A | 0.260 (0.165) | 0.250 | -0.028 | 0.164 | 0.353 | 0.645 |
| Trial B | 0.342 (0.201) | 0.372 | -0.157 | 0.239 | 0.450 | 0.721 |

Summary statistics of individual $r_s$ values for the relationship between confidence decisions and matching face pair difficulty are shown for trainees and controls in Table 26. For trial A trainee confidence decisions were more strongly correlated with match pair difficultly than controls. On trial B both groups showed an increase in $r_s$ values, indicating that on retest both group's confidence decisions were better calibrated with the difficulty of matching face pairs. However, for non-matching facial pairs $r_s$ values did not increase to the same extent on trial B for either group (see Table 27).

*Table 27 - Summary statistics of Spearman's rho values of confidence-accuracy relationship for non-match pairs*

| Correlation coefficients for non-match pairs (Spearman's Rho) | | | | | | |
|---|---|---|---|---|---|---|
| Trainee | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Trial A | 0.344 (0.167) | 0.404 | -0.019 | 0.228 | 0.450 | 0.562 |
| Trial B | 0.385 (0.147) | 0.400 | 0.068 | 0.315 | 0.520 | 0.575 |
| | | | | | | |
| Control | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Trial A | 0.339 (0.165) | 0.350 | 0.092 | 0.196 | 0.470 | 0.652 |
| Trial B | 0.322 (0.195) | 0.330 | -0.061 | 0.190 | 0.498 | 0.684 |

The distributions of individual $r_s$ values for matching pairs and non-matching pairs are shown in Figure 21 and Figure 22 respectively. Both groups have a wide range of $r_s$ values, with some participants confidence decisions having no correlation with item difficulty and others a strong positive correlation. This indicates that individuals are using confidence ratings in very different ways.

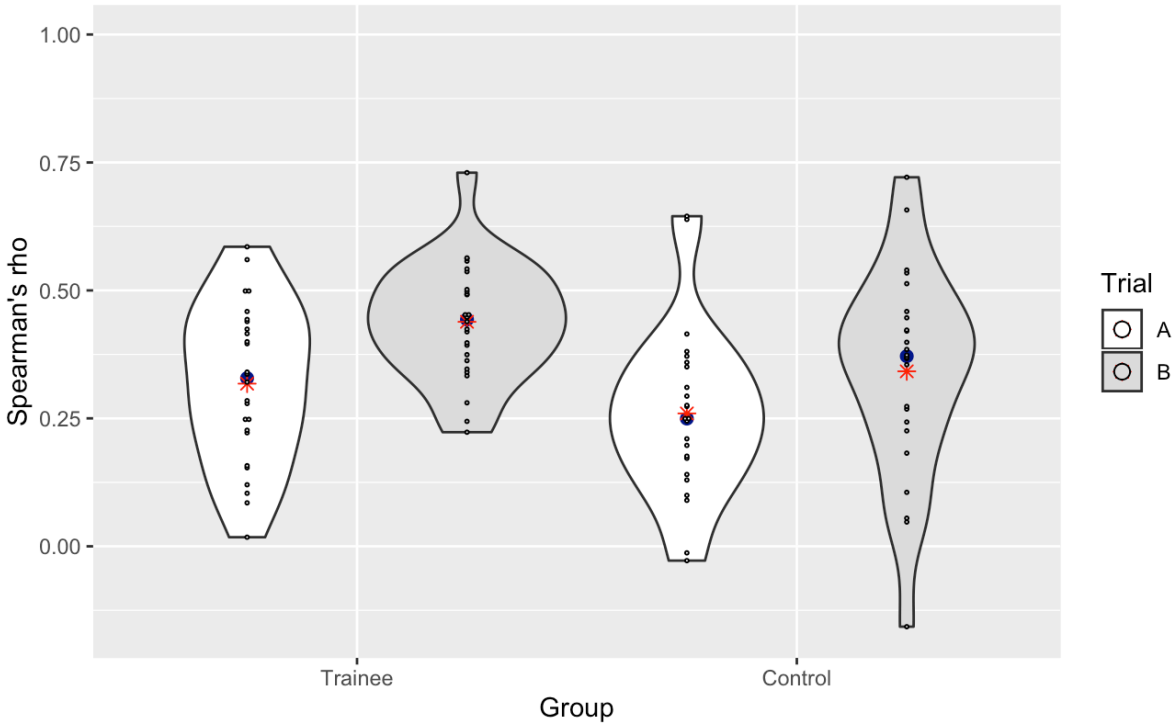


*Figure 21 – Distribution of Spearman's rho values of confidence-accuracy relationship for match pairs by group and trial (blue dot represents median group score and red star is the mean group score)*

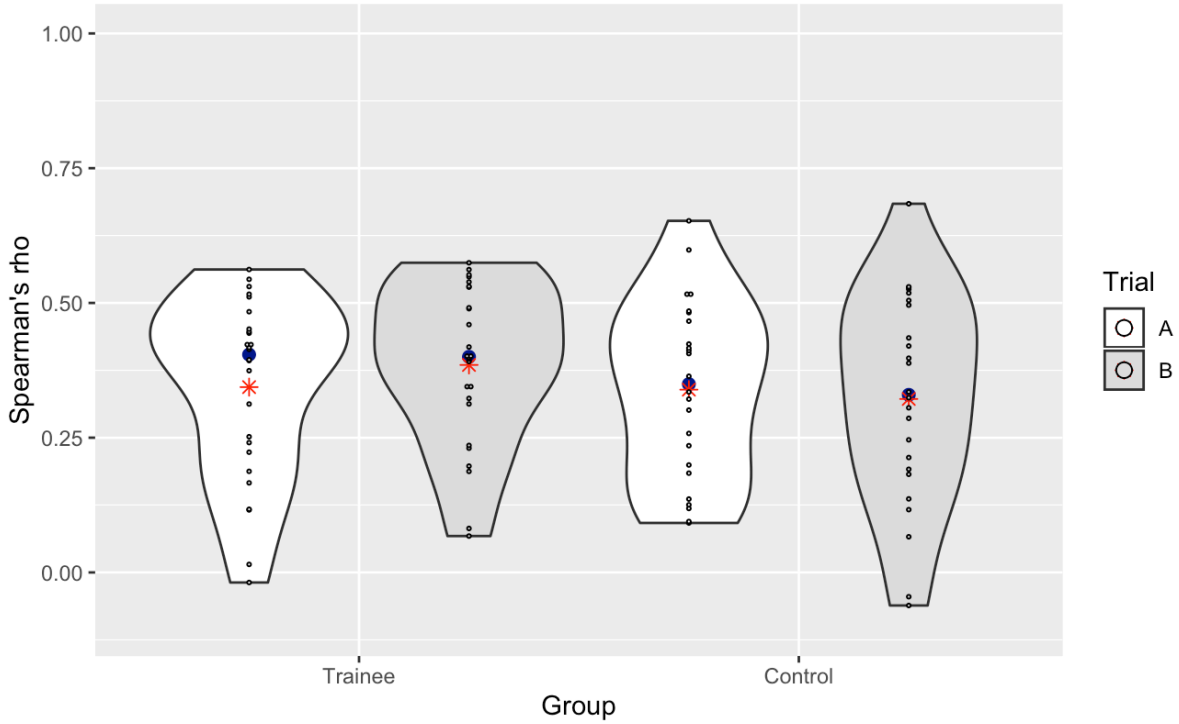


*Figure 22 – Distribution of Spearman's rho values of confidence-accuracy relationship for non-match pairs by group and trial (blue dot represents median group score and red star is the mean group score)*

A factorial ANOVA for confidence calibration on match pairs revealed a significant (albeit marginal) effect for group ($F(1,48) = 4.46$, $\eta_p^2 = .057$, $p = .040$) and a highly significant effect for trial ($F(1,48) = 14.52$, $\eta_p^2 = .095$, $p < .001$). No interaction was found between group and trial ($F(1,48) = 0.52$, $\eta_p^2 = .003$, $p = .472$). Post hoc pairwise tests for group revealed a significant difference between trainees and controls on trial B with a medium effect size ($t(87.6) = -2.13$, $d = 0.60$, $p = .036$). Post hoc tests for trial type revealed a significant difference for controls between trial A and trial B, with a medium effect size ($t(48) = -2.14$, $d = 0.45$, $p = .037$) and a significant difference for trainees between trial A and trial B with a large effect size ($t(48) = -3.27$, $d = 0.90$, $p = .002$). These results indicate that on trial B both trainees and controls confidence decisions were better calibrated to the difficulty of matching face pairs. For trainees, the effect size of the improvement was larger, which may

be caused by training. However, given that the control group also improved this could be caused by a practice effect from trial A.

For confidence calibration on non-match pairs a factorial ANOVA revealed no effect for group ($F(1,48) = 0.73$, $\eta_p^2 = .010$, $p = .396$) or trial ($F(1,48) = 0.20$, $\eta_p^2 = .001$, $p = .658$) and no interaction between trial or group ($F(1,48) = 1.18$, $\eta_p^2 = .007$, $p = .282$). In contrast to match pairs there was no reliable improvement in confidence calibration for non-match pairs on trial B for either group.

## *5.4. Discussion*

The current study found that a short professional face-matching training course produced no improvements in the overall face-matching accuracy of trainees. The training course in this study closely resembled a course evaluated by Towler et al. (2019) in a previous study (course D). Like Towler at al., this study did find some improvement in accuracy post training on images from the MFMT, but a significant improvement in accuracy was also observed for controls who did not receive any training. Given that controls also improved in accuracy on MFMT images from trial B, training does not appear to be driving this improvement. It may be that there is a practice effect causing the improvement in control accuracy on trial B or an issue with the counterbalancing of difficulty between trial A and trial B for these particular images. There was no increase in accuracy for images from the EFCT or the casework test (CWT), replicating findings from Towler et al. Unlike Towler et al., no significant improvement in accuracy was observed for images from the GFMT post training. This demonstrates not only that the improvements in face-matching accuracy from short training courses are limited, but where improvements are observed they are inconsistent and hard to replicate.

Although there was no overall improvement in face matching accuracy, some evidence was found for changes in performance on matching face pairs, but only for trainees that were initially poor at face matching. The 10 trainees that performed most poorly prior to training showed a significant increase in accuracy on matching face pairs. This increase was not observed for top performing trainees or controls and there was no reliable improvement in accuracy for non-matching face pairs in either group. There are two possible reasons for the change in match accuracy for initially poorer performing trainees. The first is that the training course genuinely improved the low performers ability to detect matches. The second is that the training course caused a shift in response bias, where the trainees respond match more frequently after training but were not more adept at detecting matches.

The training course evaluated in the study aimed to teach a feature-by-feature approach to face matching, which is believed to differ from the holistic and configural processes we naturally use to match faces (Towler et al., 2021). Previous studies have observed improvements in accuracy for matching face pairs but not non-matching pairs when novices are instructed to use a feature-based face-matching approach (see Megreya & Bindemann, 2018; Towler, White, et al., 2017), so it is perhaps unsurprising that an effect on trainee match accuracy was observed in the current study. But if this is the case, it *is* surprising that the effect was only observed for trainees who were initially poor at face matching to begin with. Analysis using signal detection measures revealed that the low performing trainees had a significant shift in response bias after training, meaning these participants were more likely to respond match as a result of training. There was also an increase in sensitivity, but this was not significant compared to pre-training and still far below the sensitivity of high performers. This suggests that the increase in match accuracy for low trainees is, in large part, derived from a shift in response bias.

It is not readily apparent why this shift in response bias from training only affects low performers. When providing examples of labelled matching and non-matching stimuli during a face-matching task Gentry & Bindemann (2019) observed an improvement in the accuracy of low performers, which they attribute to stabilisation in response criterion caused by the examples. They further postulate that the effect is only seen for low performers because they have a less stable response criterion than high performers to begin with. Based on the findings of Gentry & Bindemann, it may be that the less-stable decision criteria of low performers were more susceptible to change by training strategies. As there is evidence that the feature-based strategy taught in the current course appears to favour accuracy on match pairs, this training approach could be having a greater impact on the less-stable criterion of the low performers resulting in a shift to a more liberal criterion for this group. A note of caution is given regarding this interpretation. Given that other findings from Towler et al. (2019) were not replicated in this study, despite the similarities in experimental design, it appears that effects from short training courses can be hard to replicate. Therefore, further validation is required to confirm whether a shift in the response bias of low performers is a general effect of face-matching training, or a phenomenon unique to this study.

If a shift towards a more liberal response for low performers is a general effect of short face-matching training courses, an important question is whether this effect is advantageous or not in operational settings. Because of the shift in response bias, low performers are more likely to detect hits or matches after training, just by the virtue of responding match more frequently. However, a change in response bias also comes with the potential risk of an increased false-alarm rate (i.e. responding match to non-matching faces). False alarms may be acceptable if the consequences of making a such an error are negligible. In which case the training effect in this study would be advantageous. However, in many applied settings

the consequences of a false alarm can be severe and potentially life changing, from wrongful arrest in a police investigation to failing to detect a fraudulent non-matching passport at the border. Therefore, an effective training course should not only increase the detection of matches but also lead to a decrease in the false-alarm rate, which was not observed in this study. The base-rate probabilities of match and non-match occurrence are also factors to consider when evaluating the significance of a change in sensitivity and response bias. In this study matches and non-matches occurred with approximately equal frequency in the test stimuli, however in applied settings this is not always the case. For example, at the border non-matches are highly infrequent in relation to the prevalence of matches. This can result in a phenomenon known as the low-prevalence effect, where infrequent targets are much more likely to be missed (Papesh et al., 2018). In the border scenario, a training course that introduces a liberal response bias means trainees will be more likely to respond match to non-matching faces, potentially exacerbating the low-prevalence effect. This could inflate the false-alarm rate to a greater extent and increase the risk that infrequent non-matches are missed.

Finally, the study examined the impact of training on face-matching confidence decisions. A hallmark of forensic examiner expertise is not making high-confidence errors, and this hallmark is thought to be derived from examiner training and experience (Towler et al., 2018). In this study there was no consistent change in confidence decisions post training, demonstrating that the short training course had no obvious effect on confidence. For match trials confidence decisions were better calibrated with the difficulty of a face pair on trial B, but this was observed for both controls and trainees. Although the effect for trainees was greater, this effect cannot be attributed solely to training because of the increase in the control group. Instead the increased calibration of confidence for match pairs may be caused by repeated practice. However, feedback on decisions was not provided to the

control group so it is not clear how a practice effect could have manifested. This observation requires further investigation, particularly as the effect was not observed for non-matching face pairs

This study further demonstrates the limitations of using only short professional courses to train face-matching operators. The results found no reliable improvements in overall face-matching accuracy and little evidence for training causing changes in confidence decisions. The training did appear to increase low performers accuracy on matching face pairs, but as there was no significant improvement in sensitivity, this effect appears to be driven by a shift in response bias rather than improved discrimination . This finding highlights the need to consider the impact of training strategies on both matching and non-matching face pairs. When designing training strategies, consideration should also be given to the base-rate probabilities of match and non-match occurrence, as shifts in response bias caused by training could have dramatically different effects depending on the frequency of occurrence for matching and non-matching faces. There is now a growing body of empirical evidence that short training courses do not result in the development of reliable, superior face-matching expertise. Given the societal importance of ensuring that face-matching operators are reliable and accurate, the research and applied communities must come to together to develop training strategies and approaches to applied face-matching that are evidence-based and proven to result in more accurate face-matching decisions.

# 6. Study Three – Comparing perceptual skill and crowd effects for superior face matchers and face examiners

## 6.1. Introduction

Facial reviewers are face-matching professionals that work in high-throughput settings, making large numbers of face-matching decisions often with automated facial recognition technology. Facial reviewers are relied upon to make face-matching decisions in situations where errors can have far-reaching consequences. An international survey of face-matching training practices has revealed that several agencies use short professional training courses of five days or less to train facial reviewers (Chapter 4). In Chapter 5, the evaluation of a two-day face-matching training course demonstrated that there was no overall improvement in trainee accuracy after completion of the course, which replicates findings from previous studies (Towler et al., 2019; Woodhead et al., 1979). The limited efficacy and use of short training courses may provide a possible explanation for why, in a number studies, facial reviewers show no evidence of superior face-matching ability in standardised tests (see White et al., 2021).

The limited effectiveness of short training courses presents a challenging dilemma for how agencies should recruit and deploy facial reviewers. To tackle this challenge both the research and applied face-matching communities must come together to identify evidence-based practices for the selection and training of facial reviewers with superior face-matching ability, which will reduce the chance of incorrect face-matching decisions occurring in high-risk and security critical environments. The aim of this study is to investigate three possible avenues for improving face-matching accuracy in high-throughput environments.

The first possible solution is the selection and deployment of individuals with innately superior face-matching skills, commonly referred to as super recognisers (SR) or more specifically super matchers (Bate et al., 2018), into operational roles. Bobak et al. (2016) suggested that individuals with superior face memory could be recruited for face-matching roles in border control. Seven SRs, recruited through a media campaign and confirmed using pre-testing on the CFMT+, took part in two face-matching experiments. The seven SRs outperformed controls at the group level but not all SRs were significantly superior at an individual level. A potential limitation of Bobak et al.'s study is that the SRs were recruited based on their face memory skills but tested using face-matching tasks. More recent research has found that although face memory and face matching ability are correlated, there was limited evidence that high performance on one task generalised to high performance on another task (Fysh et al., 2020). Researchers, therefore, advocate the use of selection tests that are representative of the types of tasks that the face-matching operator will be required to do in their day-to-day duties (Moreton et al., 2019). Despite this recommendation, the survey results presented in Chapter 4 show that only half of the responding agencies pre-screened facial reviewers prior to training, highlighting a possible disconnect between research and practice.

The second solution under investigation is to use face examiners to improve the accuracy of face-matching decisions in high-throughput environments. Forensic face examiners have consistently demonstrated enhanced face-matching accuracy at the group level (White et al., 2021). Face examiners predominantly match faces using a detailed, morphological analysis approach, which can take several hours or even days per case depending on the complexity of the task (Moreton, 2021). Therefore, the morphological feature-based approach used by examiners is not applicable in high-throughput face-matching environments, such as checking passports at the border, where operators must potentially

make a decision in seconds (Stevens, 2021). However, face examiners have also shown enhanced perceptual skills in quick decision face-matching tasks when decisions are made in under 30 seconds (White, Phillips, et al., 2015). According to the survey results from Chapter 4 only a minority of agencies pre-screen examiners prior to training. Most examiners are trained for one-to-five years, so there is clearly a difference in reviewer and examiner training practices. But based on current research it is not clear whether this enhanced perceptual skill is derived from training or if examiners are naturally superior face-matchers and thus attracted to these roles. The number of face examiners employed within different agencies are not known, but they are understood to operate in small, specialist units. The combination of lengthy training and small numbers means examiners, whilst performing well at quick, perceptual face-matching tasks, are unlikely to be a sustainable and scalable solution for high volume, quick decision face matching in applied settings.

The third solution investigated in this study is the wisdom of crowds. The crowd sourcing of face-matching decisions has been shown to consistently improve performance (Jeckeln et al., 2018; White et al., 2013), and is particularly effective when combining decisions from multiple SRs and face examiners (Phillips et al., 2018).

This study evaluated and compares the perceptual face-matching of two groups of high performers on a face-matching task and then explored the wisdom of crowds as a means to improve accuracy. Firstly, superior face matchers were selected from a pool of police officers and staff using a challenging face-matching task. These superior face-matchers were then tested on a second face-matching task that closely resembled the selection task. The performance of the superior face matchers at re-test was compared to controls and a small group of trained forensic face examiners. The face examiners were not allowed to use their standard tools and procedures to complete the task, in order to test their perceptual skill in face matching rather than their facial examination strategies. Finally, different

iterations of face-matching crowds were created and evaluated using individuals from the high performing groups. Results were analysed at the group level and using individual case analysis, as recommended by Bobak, Hancock, et al. (2016).

## *6.2. Methods*

### 6.2.1. Participants

A control group of 138 police officers and staff from a UK police force (39 female) and three trained forensic face examiners (2 female, with professional face-matching experience ranging from 12 to 84 months) participated in the first half of the study, by completing a face-matching task (trial A). Due to high levels of attrition, of the 138 control participants, only 28 completed a follow up face-matching task (trial B). These 28 controls who completed both trial A and trial were used as a selection pool for high performing face matchers with superior perceptual skill. The three forensic face examiners also completed trial B.

### 6.2.2. Materials

Trial A, consisting of 107 face pairs (54 matching pairs and 53 non-matching pairs). Trial B consisted of 108 face pairs (55 matching pairs and 53 non-matching pairs). These are the same materials used in Chapter 5.

### 6.2.3. Procedure

Participants completed face-matching trial A and face-matching trial B on different days following the same procedure used in Chapter 5 for control participants. Prior to completing the trials participants consented to take part in the study. The study received a favourable opinion from the ethical committee of the Open University.

## *6.3. Results*

### 6.3.1. Trial A short form and Trial B short form

Tests of perceptual skill must be suitably challenging to identify individuals with superior performance. If the items in a test are too easy there may be ceiling effects and the test will not be sufficiently sensitive to find truly superior individuals. To overcome this potential limitation trial A and trial B were both reduced to their 50 most challenging face pairs, referred to as trial A short form and trial b short form. Trial A short form and trial B short form were used to evaluate the perceptual skill of superior face matchers and forensic face examiners

Item-based analyses of trial A and trial B were undertaken to understand the distribution of scores per face pair and thus the difficulty of different face pairs. Item scores were calculated as the percentage of correct decisions for an image pair made by untrained control participants (N = 138 for trial A, N = 58 for trial B[10]). The distribution of item scores revealed 36%of face pairs in trial A and 34% of face pairs in trial B received correct responses over 90% of the time (see Figure 23 and Figure 25). These items were the least challenging face pairs and of limited usefulness in identifying and testing superior face matchers. Using the item difficulty scores from trial A and trial B, the 25 hardest matching and 25 hardest non-matching face pairs were selected to form short form versions of each trial. The distribution of item scores for trial A and trial A short form are shown in Figure 23 and Figure 24 and summary statistics comparing item difficulty between the two tests are shown in Table 28. The summary statistics demonstrate that trial A short form contains

[10] Differences in N between trial A and trial B are due to drop out rates of police controls. Only 28 participants completed both trial A and B, who then became the selection pool for superior face matchers.

more challenging face matching pairs. In both forms of trial A the non-matching pairs are, overall, more challenging than the matching pairs but both show a wide range in difficulty. Control performance on trial A and trial A short form were, as expected, strongly correlated, $r(138) = .93$, $p < .001$.

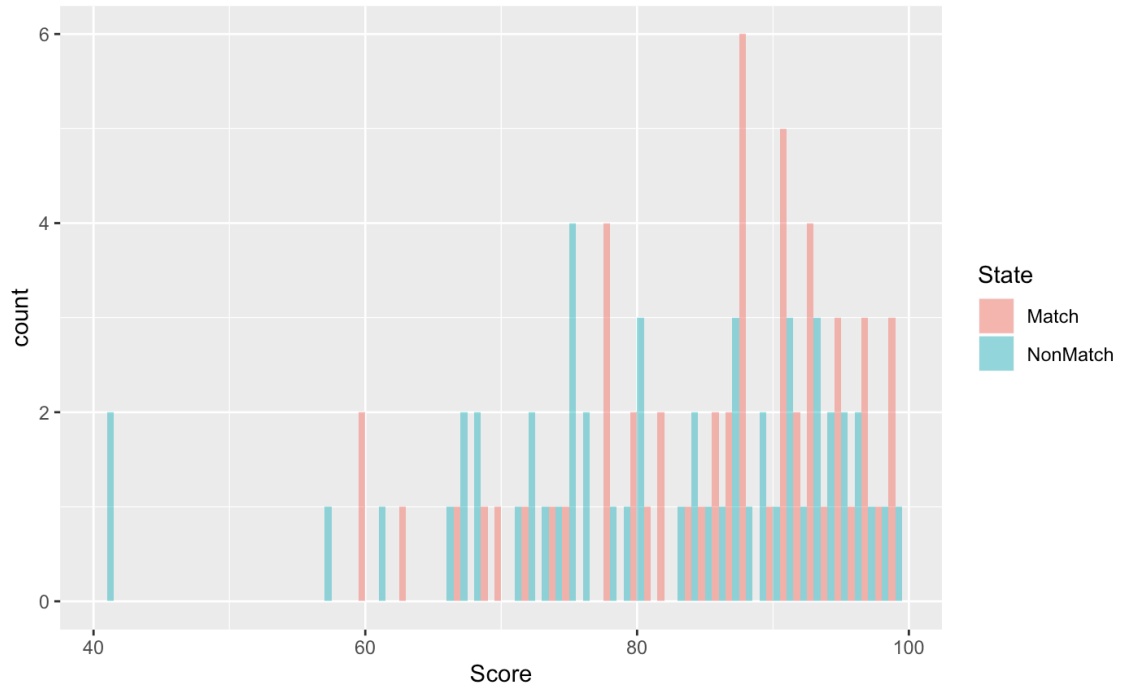


*Figure 23 – Distribution of scores for all face image pairs from trial A*

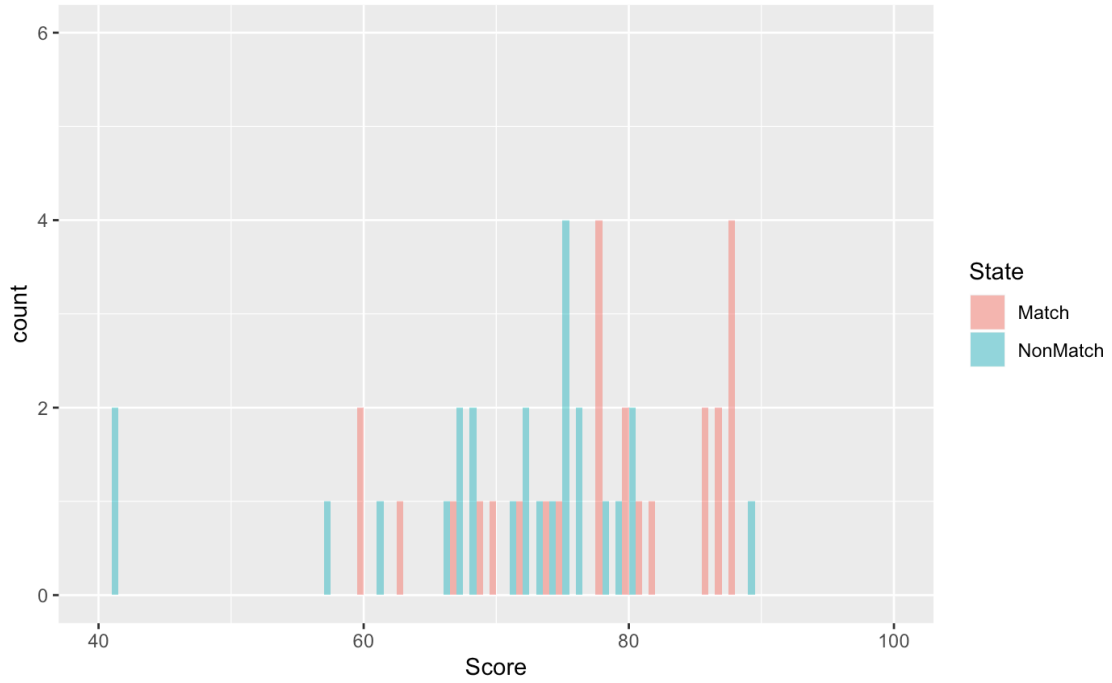


*Figure 24 – Distribution of scores for 25 hardest matching and 25 hardest non-matching face image pairs from trial A*

*Table 28 - Summary statistics of item difficulty for Trial A long form and Trial A short form*

| Item difficulty | | | | | | |
|---|---|---|---|---|---|---|
| Trial A long form | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Overall | 83.41 (11.98) | 86.96 | 40.58 | 75.72 | 92.75 | 98.55 |
| Match pairs | 85.78 (10.17) | 87.68 | 60.14 | 80.43 | 93.48 | 98.55 |
| Non match pairs | 81.00 (13.24) | 84.06 | 40.58 | 73.91 | 91.30 | 98.55 |
| | | | | | | |
| Trial A short form | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Overall | 73.99 (10.68) | 75.00 | 40.58 | 68.30 | 80.43 | 89.13 |
| Matches pairs | 77.74 (8.99) | 78.26 | 60.14 | 72.46 | 86.23 | 88.41 |
| Non matches pairs | 70.23 (11.09) | 73.19 | 40.58 | 66.67 | 76.09 | 89.13 |

The distributions of scores for Trial B and Trial B short form are shown in Figure 25 and Figure 26 with summary statistics shown in Table 29. Trial B short form contains more challenging face matching pairs and is of a similar difficulty to trial A short form. As for trial

A short form, the non-matching face pairs in trial B short form are more challenging than the matching face pairs. Performance on both forms of trial B was strongly correlated, $r(58) = .94$, $p < .001$.

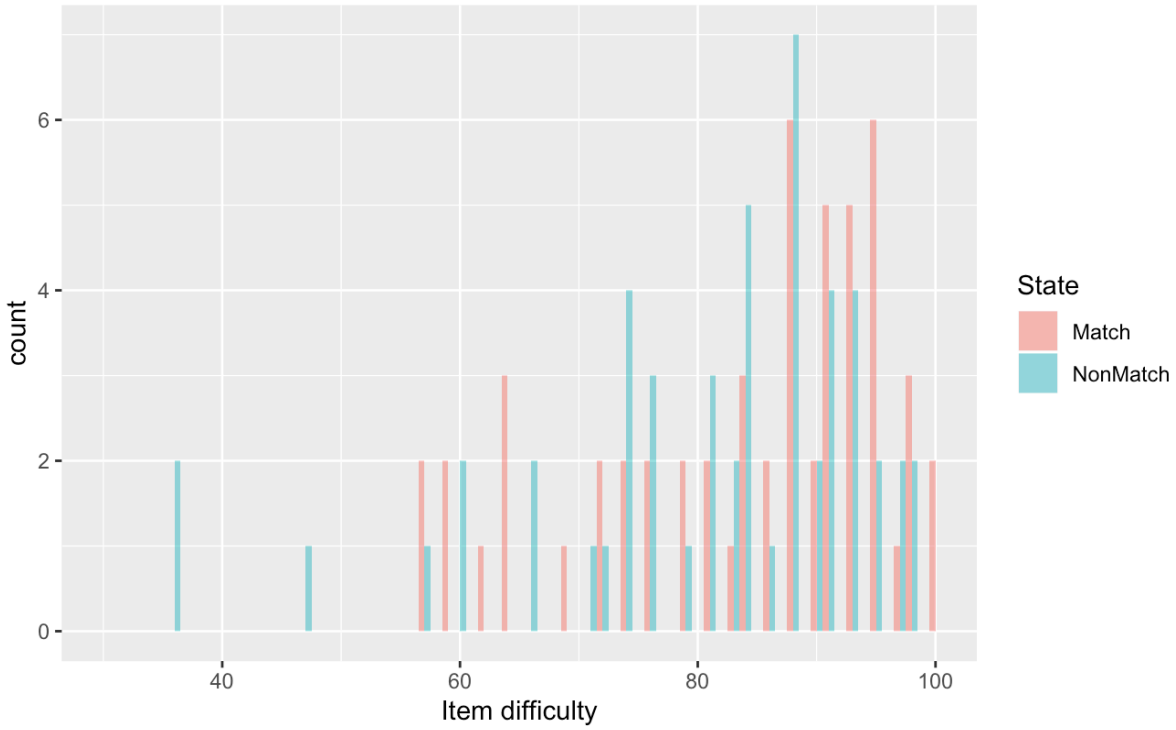


*Figure 25 – Distribution of scores for all face image pairs from trial B*

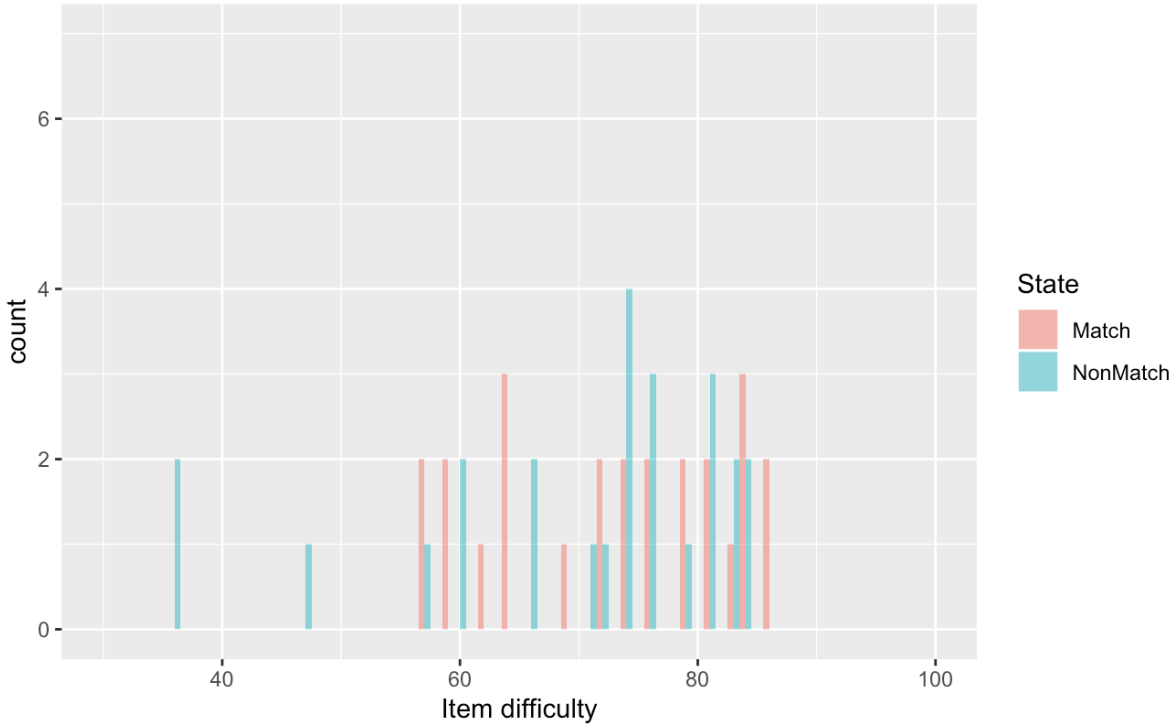


*Figure 26 – Distribution of scores for 25 hardest matching and 25 hardest non-matching face image pairs from trial B*

*Table 29 – Summary statistics of item difficulty for Trial B long form and Trial B short form*

| Item difficulty | | | | | | |
|---|---|---|---|---|---|---|
| Trial B long form | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Overall | 82.60 (13.46) | 87.07 | 36.21 | 75.43 | 93.10 | 100.00 |
| Match | 83.95 (12.31) | 87.93 | 56.90 | 75.86 | 93.10 | 100.00 |
| Non match | 81.20 (14.53) | 84.48 | 36.21 | 74.14 | 91.38 | 100.00 |

| Trial B short form | | | | | | |
|---|---|---|---|---|---|---|
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Overall | 71.59 (12.05) | 74.14 | 36.21 | 63.79 | 81.03 | 86.21 |
| Matches | 73.10 (9.98) | 74.14 | 56.90 | 63.79 | 81.03 | 86.21 |
| Non matches | 70.07 (13.85) | 74.14 | 36.21 | 65.52 | 81.03 | 84.48 |

A correlation analysis of accuracy on trial A short form and trial B short form revealed a strong positive relationship, $r(31) = .74$, $p < .001$, demonstrating that performance on one task was predictive of performance on the other.

### 6.3.2. Selecting superior face matchers

Due to a high dropout rate only 28 individuals from the sample of 138 controls completed both trial A and trial B. These 28 individuals formed a selection pool for identifying superior face matchers, where their face-matching performance could be evaluated across two related face-matching tasks (trial A short form and trial B short form). Superior face matchers were identified as the top performers from the selection pool based on their accuracy on trial A short form, simulating how top performers may be identified using face perception tests in operational settings.

Descriptive statistics were used to compare the face-matching ability of the selection pool to the larger 138 participant control group on trial A short form. This was to ensure that the selection pool was representative of the range of face-matching abilities observed in the larger control group and not skewed towards only high or low performers. Table 30 shows summary statistics of accuracy for the selection pool and larger control group on trial A short form. Variance in accuracy between the selection pool and the larger control group appeared to be equivalent, confirmed with Levene's test for homogeneity of variance ($F(1,164) = 0.08$, $p = .776$), indicating that the selection pool were representative of the overall variance in face-matching accuracy within the larger control group.

*Table 30 – Summary statistics of control and selection pool accuracy on trial A short form*

| Trial A short form | | | | | | |
|---|---|---|---|---|---|---|
| Control accuracy | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Overall | 73.99 (10.51) | 74.00 | 38.00 | 66.00 | 82.00 | 96.00 |
| Match | 77.74 (16.62) | 80.00 | 28.00 | 69.00 | 88.00 | 100.00 |
| Non match | 70.23 (17.37) | 72.00 | 28.00 | 60.00 | 84.00 | 100.00 |

| Selection pool accuracy | | | | | | |
|---|---|---|---|---|---|---|
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Overall | 73.93 (10.21) | 74.00 | 54.00 | 65.50 | 82.00 | 94.00 |
| Matches | 75.86 (17.60) | 76.00 | 28.00 | 67.00 | 89.00 | 100.00 |
| Non matches | 72.00 (17.32) | 72.00 | 32.00 | 60.00 | 85.00 | 100.00 |

Figure 27 shows the distribution of accuracy on trial A short form for the control and selection pool subset. A Shapiro-Wilk test confirmed that accuracy for the larger control group was normally distributed (W = 0.983, p = 0.087), therefore the mean and standard deviation of this group were appropriate measures for identifying superior face matchers.

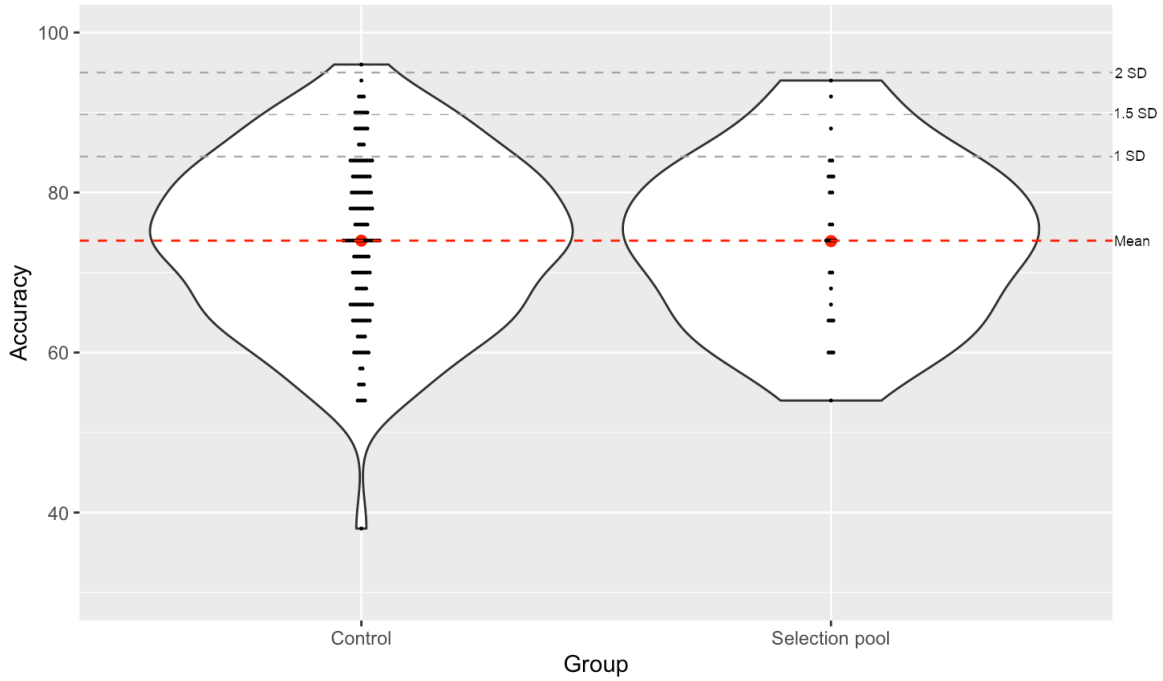


*Figure 27 - Accuracy for trial A short form by group with control mean, 1 SD, 1.5 SD and 2 SD cut offs*

Selection cut-offs for superior face-matching accuracy on trial A short form were established using the mean accuracy and standard deviation of the larger control group, which are shown in Figure 27. Within the selection pool three individuals had an accuracy score greater than one standard above the mean and two of these individuals were greater than one and a half standard deviations above the mean. No individual within the selection pool achieved an accuracy score greater than two standard deviations above the mean, which is the recommended selection criteria for super recognisers on standardised tests (Bobak, Pampoulov, et al., 2016). The top three performers from the selection pool were identified as superior face-matchers.

Figure 28 shows the distribution of match and non-match scores on trial A short form for the selection pool and three selected superior face-matchers (SMs). Mean accuracy and one standard deviation cut offs are derived from the larger 138 participant control group. There is variation in accuracy on matching and non-matching pairs between the three SMs. One of the SMs performed one standard deviation above the mean for both matching and non-matching pairs, another SMs was at ceiling for non-matching pairs but did not perform as well for matching pairs. The final SM performed one standard deviation above the mean for matching pairs but not for non-matching pairs. This demonstrates that the three, high performing SMs are not homogenous in their face-matching decision making.

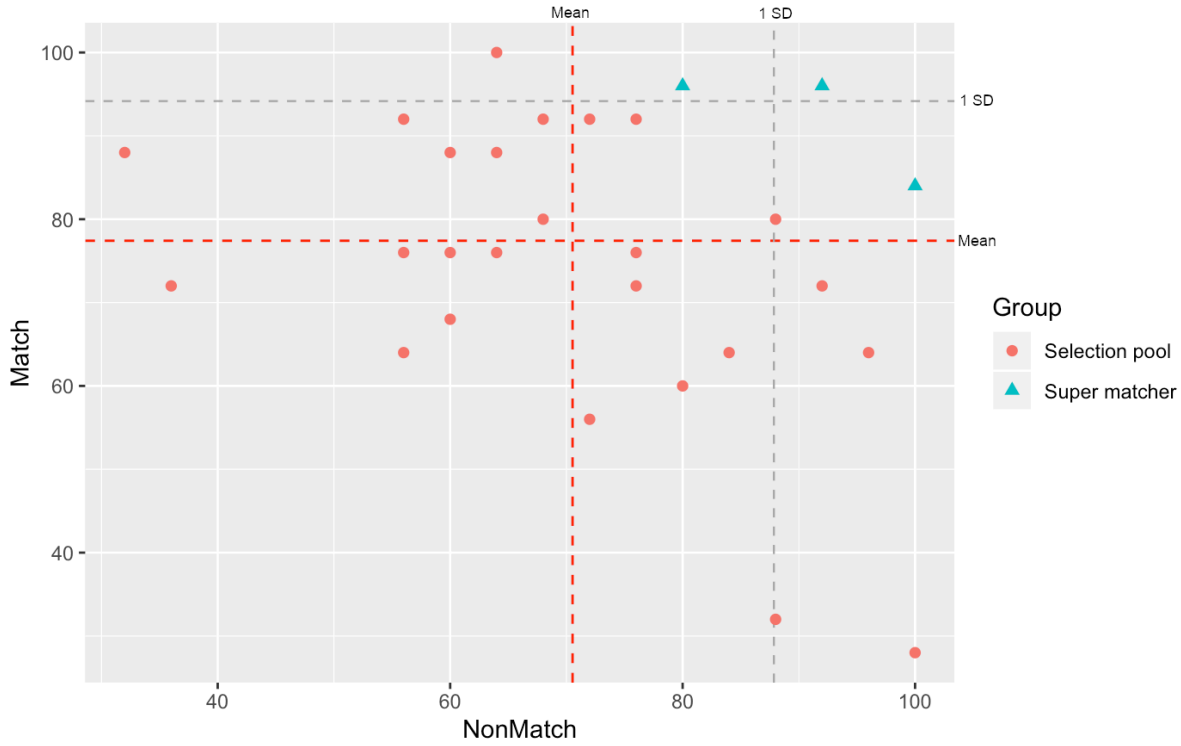


*Figure 28 – Match and non-match accuracy of superior face matchers on Trial A short form*

### 6.3.3. Accuracy on trial A short form

The 138 participant control group were used in group level and individual case analyses to evaluate the performance of the three superior face matchers and three face examiners (FEs) on trial A short form. The performance of each group was compared in terms of overall accuracy, match accuracy and non-match accuracy, with summary statistics for each group shown in Table 31. Both SMs and FEs outperformed the control group in terms of overall accuracy, with similar levels of performance between SM and FE groups. For FEs accuracy on match trials was slightly greater than that of SMs whereas non-match accuracy was similar between the two groups.

*Table 31 – Summary statistics of overall, match and non-match accuracy for Trial A short form by group*

| Trial A short form | | | | | | |
|---|---|---|---|---|---|---|
| Controls (N = 138) | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Overall | 73.99 (10.51) | 74.00 | 38.00 | 66.00 | 82.00 | 96.00 |
| Matches | 77.74 (16.62) | 80.00 | 28.00 | 69.00 | 88.00 | 100.00 |
| Non matches | 70.23 (17.37) | 72.00 | 28.00 | 60.00 | 84.00 | 100.00 |
| SMs (N = 3) | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Overall | 91.33 (3.06) | 92.00 | 88.00 | 90.00 | 93.00 | 94.00 |
| Matches | 92.00 (6.93) | 96.00 | 84.00 | 90.00 | 96.00 | 96.00 |
| Non matches | 90.67 (10.07) | 92.00 | 80.00 | 86.00 | 96.00 | 100.00 |
| FEs (N = 3) | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Overall | 92.67 (5.03) | 92.00 | 88.00 | 90.00 | 95.00 | 98.00 |
| Matches | 96.00 (4.00) | 96.00 | 92.00 | 94.00 | 98.00 | 100.00 |
| Non matches | 89.33 (6.11) | 88.00 | 84.00 | 86.00 | 92.00 | 96.00 |

A one-way ANOVA revealed a significant difference in overall accuracy between the groups ($F(2,140) = 9.14$, $\eta_p^2 = .116$, $p < .001$). Post hoc pairwise tests with Bonferroni adjustment revealed a significant difference between FEs and controls with a large effect size ($t(140) = -3.17$, $d = 1.84$, $p = .006$). The difference between SMs and controls was also significant, however this was to be expected as they were selected as the top performers from a subset of controls on this trial ($t(148) = -2.94$, $d = 1.71$, $p = .012$). No significant difference was found between SMs and FEs in overall accuracy ($t(140) = 0.16$, $d = 0.32$, $p = 1$). The distributions of overall accuracy for each group are shown in Figure 29, demonstrating that there is a range in individual accuracy for both high-performing groups.

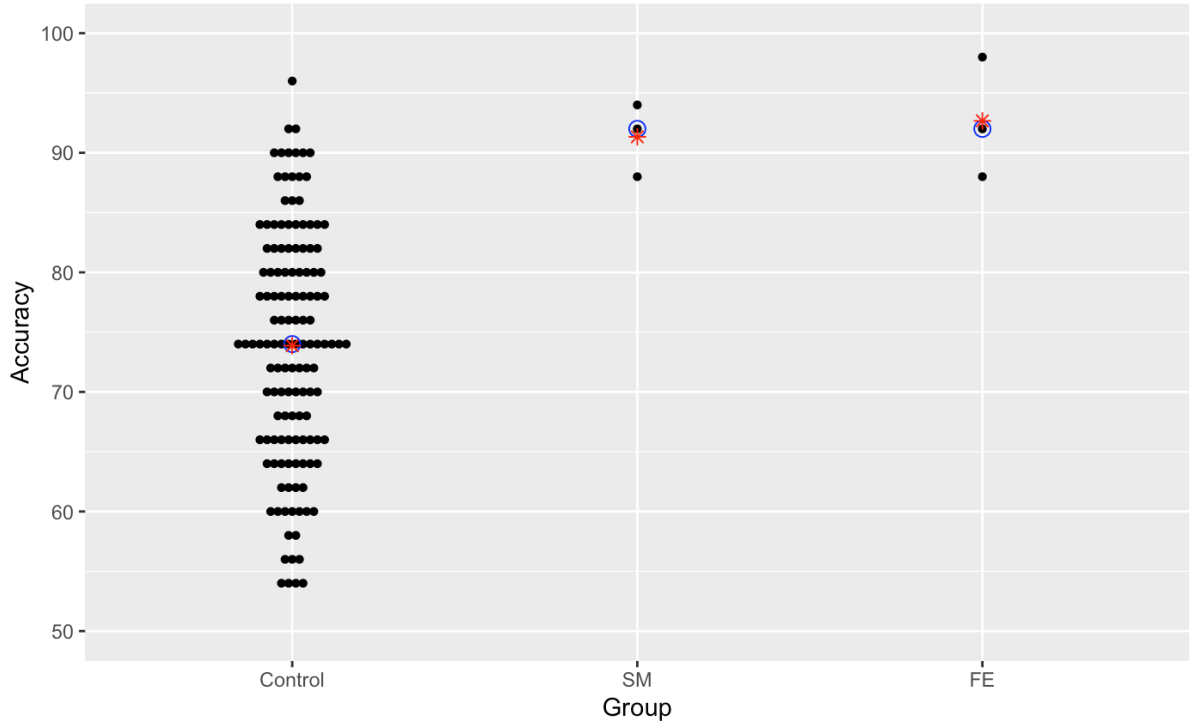


*Figure 29 – Accuracy on trial A short form by group (blue circle represents median group score and red star is the mean group score)*

One-way ANOVAs revealed no significant difference in match accuracy ($F(2,28) = 1.80$, $\eta_p^2 = .114$, $p = .183$) and no significant difference in non-match accuracy ($F(2,28) = 2.43$, $\eta_p^2 = .148$, $p = .106$) between the three groups. A scatterplot of match and non-match accuracy, shown in Figure 30, reveals that different individual controls were at ceiling for match and non-match accuracy but not for overall accuracy. This explains why a significant difference between controls and the high performing groups was found for overall accuracy but not for match and non-match accuracy.

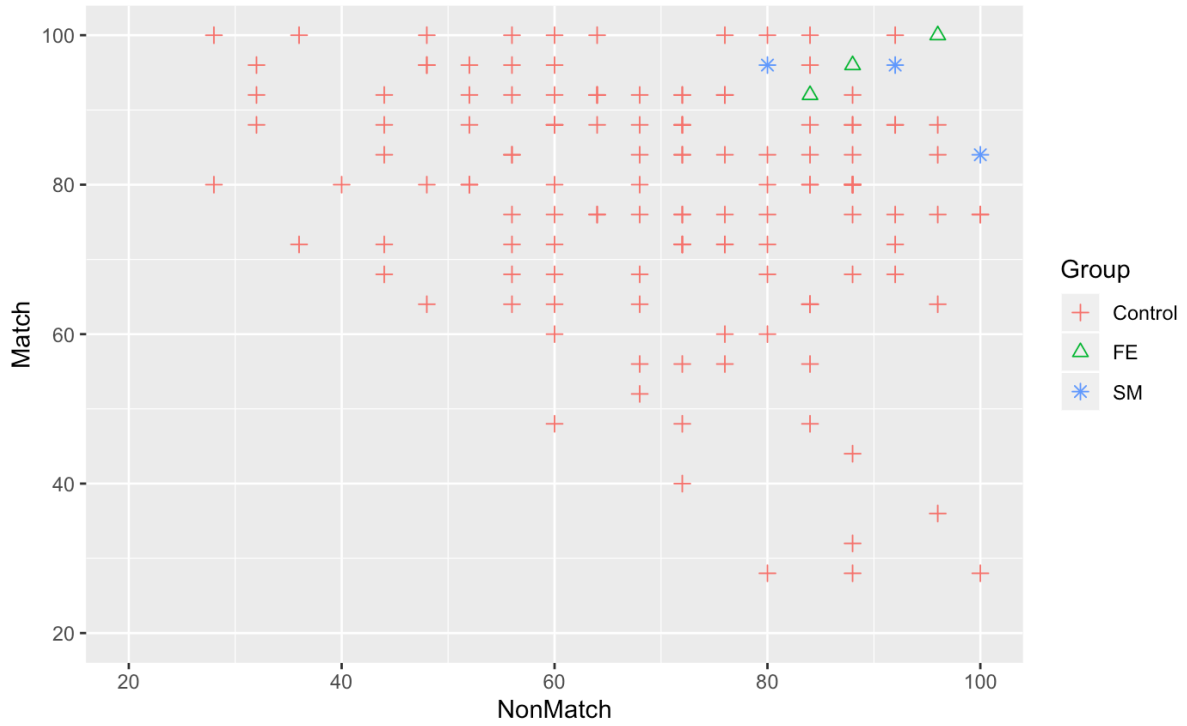


*Figure 30 – Scatterplot of match and non-match accuracy by group on trial A short form*

The next stage of the analysis used modified single-case t-tests to compare the overall accuracy of individual high-performers to the control group (Crawford et al., 2010). The aim of this analysis was to understand whether the advantage in group accuracy for SMs and FEs extended to individual high performers. Because the SMs and FEs are expected to be superior to controls in terms of overall face-matching accuracy, as demonstrated by the group-level analysis, a one-tailed test was deemed appropriate. However, for completeness, significance values for both one-tailed and two-tailed tests are reported. Individual case analyses of SMs (Table 32) revealed that all three individuals outperformed the average control score. However, only SM2 and SM3 outperformed controls at a statistically significant level. For SM2 this was for both one-tailed and two-tailed tests. SM3 outperformed controls at a statistically significant level only for the one-tailed test.

*Table 32 – Individual case analyses comparing accuracy of superior face matchers with mean control accuracy on trial A short form*

| | Mean overall accuracy (SD) | SM1 | SM2 | SM3 |
|---|---|---|---|---|
| Overall accuracy | - | 92 | 94 | 88 |
| Control (N = 134) | 73.87 (9.85) | | | |
| *t* (137) | - | 1.84 | 2.04 | 1.43 |
| p (one-tailed) | - | .035 | .021 | .078 |
| p (two-tailed) | - | .069 | .045 | .156 |
| 95% CI | - | [94.07, 98.29] | [95.94, 99.04] | [88.32, 95.30] |
| Population below individual's score (%) | - | 96.55 | 97.81 | 92.24 |

Individual case analyses for FEs revealed similar variability in accuracy to the individual SMs (Table 33). FE1 performed exceptionally well with a significant difference to controls for both one-tailed and two-tailed tests. FE3 outperformed controls with a statistically significant difference but only for the one-tailed test and FE2's accuracy was not statistically superior.

*Table 33 – Individual case analyses comparing accuracy of superior face matchers with mean control accuracy on trial A short form*

| | Mean overall accuracy (SD) | FE1 | FE2 | FE3 |
|---|---|---|---|---|
| Overall accuracy | - | 98 | 88 | 92 |
| Control (N = 134) | 73.87 (9.85) | | | |
| *t* (137) | - | 2.45 | 1.43 | 1.84 |
| p (one-tailed) | - | .008 | .078 | .035 |
| p (two-tailed) | - | .016 | .156 | .069 |
| 95% CI | - | [98.25, 99.74] | [88.32, 95.30] | [94.07, 98.29] |
| Population below individual's score (%) | - | 99.20 | 92.24 | 96.55 |

### 6.3.4. Accuracy trial B short form

The performance of the three SMs, three FEs and the selection pool controls (N= 25) were compared using face-matching trial B short form. The aim of this analysis was to demonstrate whether the high performing groups still demonstrated superior face-matching ability at re-test and if individual case analyses revealed similar patterns in performance at an individual level. Summary statistics for overall, match and non-match accuracy are shown in Table 34.

*Table 34 – Summary statistics of overall, match and non-match accuracy for Trial B short form by group*

| Trial B short form | | | | | | |
|---|---|---|---|---|---|---|
| Controls | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Overall | 71.12 (11.05) | 70.00 | 54.00 | 62.00 | 80.00 | 90.00 |
| Match | 73.44 (20.36) | 80.00 | 28.00 | 64.00 | 84.00 | 100.00 |
| Non match | 68.8 (17.81) | 68.00 | 28.00 | 60.00 | 84.00 | 100.00 |
| | | | | | | |
| SMs | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Overall | 86.67 (3.06) | 86.00 | 84.00 | 85.00 | 88.00 | 90.00 |
| Matches | 85.33 (8.33) | 88.00 | 76.00 | 82.00 | 90.00 | 92.00 |
| Non matches | 88.00 (4.00) | 88.00 | 84.00 | 86.00 | 90.00 | 92.00 |
| | | | | | | |
| FEs | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Overall | 88.00 (6.93) | 84.00 | 84.00 | 84.00 | 90.00 | 96.00 |
| Matches | 93.33 (8.33) | 96.00 | 84.00 | 90.00 | 98.00 | 100.00 |
| Non matches | 82.67 (10.07) | 84.00 | 72.00 | 78.00 | 88.00 | 92.00 |

Both SMs and FEs outperformed controls in average overall accuracy. FEs had a slightly higher overall accuracy than SMs, which appears to be driven by accuracy on match trials. Whereas for non-match accuracy SMs outperformed FEs. A one-way ANOVA revealed a significant effect for accuracy between the groups ($F(2,28) = 5.86$, $\eta_p^2 = .295$, $p = .007$).

Post hoc pairwise tests with Bonferroni adjustment revealed a significant difference between FEs and controls with a large effect size (t(28) = -2.65, d = 1.57, p = .039), however the difference between SMs and controls did not reach significance (t(28) = -2.44, d = 1.46, p = .064). No significant difference was found between SMs and FEs (t(28) = 0.16, p = 1) in overall accuracy. These results demonstrate that, although the SMs still performed well, this was not to the same extent that was observed for trial A short form. FEs still retained a significant group level advantage, however, as shown by the distributions of overall accuracy for each group in Figure 31, this was largely driven by a single FE performing almost at ceiling. The remaining two FEs performed within the range of both controls and SMs.

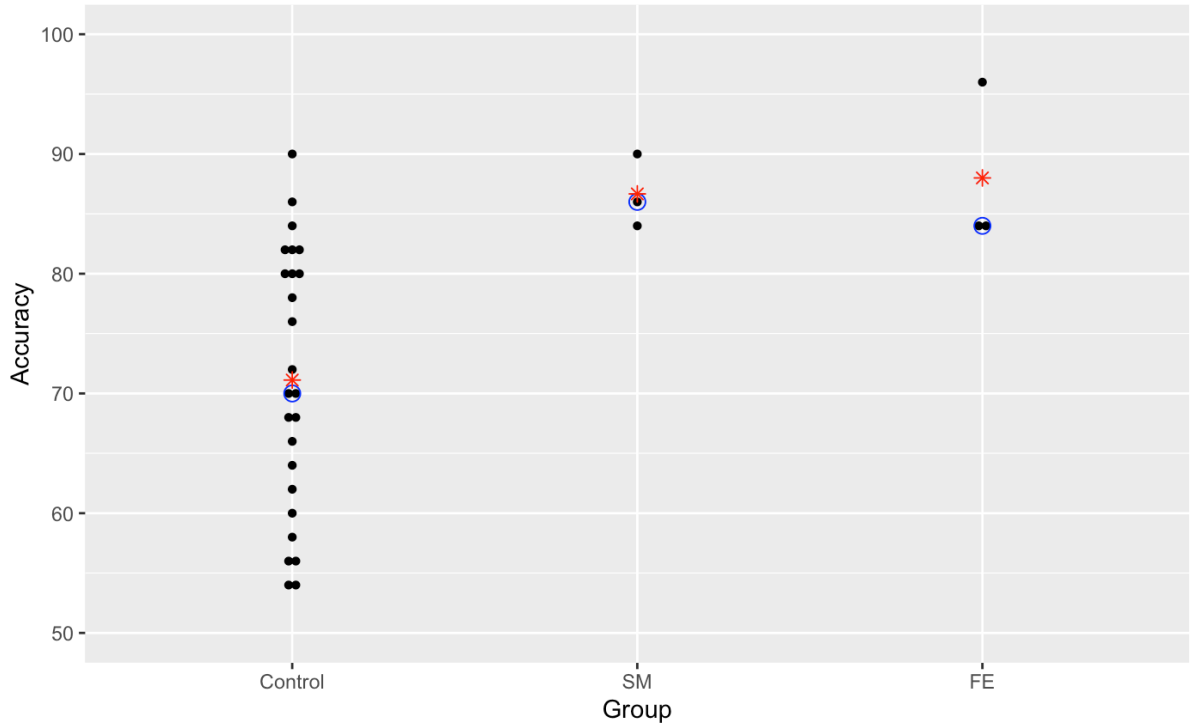


*Figure 31 – Accuracy on trial B short form by group (blue circle represents median group score and red star is the mean group score)*

As for trial A short form, one-way ANOVAs revealed no significant difference in match accuracy across the three groups ($F(2,28) = 1.80$, $\eta_p^2 = .114$, $p = .183$) and no significant difference in non-match accuracy ($F(2,28) = 2.43$, $\eta_p^2 = .148$, $p = .106$). A scatterplot of match and non-match accuracy is shown in Figure 32.

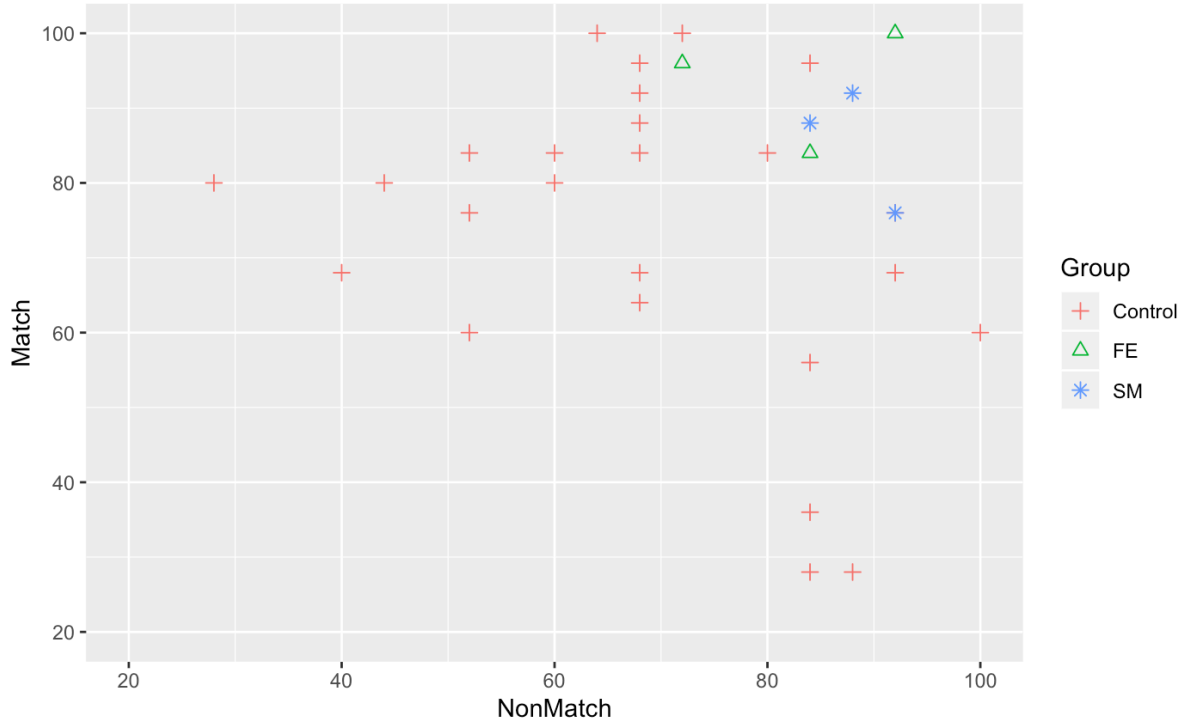


*Figure 32 – Scatterplot of match and non-match accuracy by group on trial B short form*

Modified single-case t-tests were used to compare the overall accuracy of individual SMs and FEs to the control group on trial B short form (see Table 35). Although all SMs showed an accuracy advantage compared to the control mean this was not statistically significant for any of the three individuals. Trial A short form and trial B short form were strongly correlated, therefore it is expected that high performing individuals should do well on both tests. However, as demonstrated, this was not a guarantee that superior face matchers will

perform at the same level of accuracy at re-test, even when their previous performance was greater than average at a statistically significant level.

*Table 35 – Individual case analyses comparing accuracy of superior face matchers with mean control accuracy on trial B short form*

| | Mean overall accuracy (SD) | SM1 | SM2 | SM3 |
|---|---|---|---|---|
| Overall accuracy | - | 90 | 86 | 84 |
| Control (N = 25) | 71.12 (11.05) | | | |
| *t* (24) | - | 1.71 | 1.35 | 1.17 |
| p (one-tailed) | - | 0.055 | .010 | 0.132 |
| p (two-tailed) | - | 0.107 | .200 | 0.264 |
| 95% CI | - | [86.00, 98.99] | [78.63, 97.02] | [74.10, 95.24] |
| Population below individual's score (%) | - | 94.65 | 90.04 | 86.78 |

Individual case analyses of FEs confirms that the enhanced group level accuracy on trial B short form was being driven by a single high performing individual (see Table 36). FE1 outperformed controls at a statistically significant level (p (two-tailed) = .037), with this individual retaining the accuracy advantage observed on trial A short form. Although FE2 and FE3 performed well, both were within the region of SMs and top controls and were not statistically superior at an individual level. The individual case analyses revealed that the FEs, in general, showed similar levels of perceptual skill in face matching as the SMs. FEs also showed fluctuations in performance across tests in a similar manner to SMs.

*Table 36 – Individual case analyses comparing accuracy of face examiners with mean control accuracy on trial B short form*

| | Mean overall accuracy (SD) | FE1 | FE2 | FE3 |
|---|---|---|---|---|
| Overall accuracy | - | 96 | 84 | 84 |
| Control (N = 25) | 71.12 (11.05) | | | |
| *t* (24) | - | 2.25 | 1.17 | 1.17 |
| p (one-tailed) | - | 0.019 | 0.132 | 0.132 |
| p (two-tailed) | - | 0.037 | 0.264 | 0.264 |
| 95% CI | - | [93.30, 99.86] | [74.10, 95.24] | [74.10, 95.24] |
| Population below individual's score (%) | - | 98.15 | 86.78 | 86.78 |

### 6.3.5. Sensitivity and response bias of superior face matchers and face examiners

Sensitivity and response bias were compared between trial A short form and trial B short form at the group level and for individual SMs and FEs using the Bayesian difference test developed by Crawford et al. (2011). The purpose of this analysis was to understand if sensitivity and bias were consistent for individual high performers across repeated tests, compared to the selection pool control sample (N = 25). Due to some individuals reaching ceiling on match and non-match trials and the small sizes of the SM and FE groups, non-parametric measures of sensitivity (*A*) and response bias (*b*) were used (Zhang & Mueller, 2005). For the control group both *A* ($r(25) = 0.53$, $p = .006$) and *b* ($r(25) = 0.65$, $p < .001$) were significantly correlated between trial A short form and trial B short. Summary statistics for *A* and *b* are shown in Table 37.

*Table 37 – Summary statistics of A and b for Trial A short form and Trial B short form by group*

| Control (N = 25) | | | | | | |
|---|---|---|---|---|---|---|
| Trial A short form | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| A | 0.79 (0.09) | 0.81 | 0.58 | 0.72 | 0.87 | 0.90 |
| b | -0.07 (0.52) | -0.18 | -0.82 | -0.47 | 0.26 | 1.29 |
| Trial B short form | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| A | 0.78 (0.12) | 0.78 | 0.57 | 0.70 | 0.88 | 0.94 |
| b | -0.11 (0.49) | -0.25 | -0.89 | -0.46 | 0.40 | 0.96 |
| SM (N = 3) | | | | | | |
| Trial A short form | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| A | 0.95 (0.02) | 0.95 | 0.93 | 0.94 | 0.96 | 0.97 |
| b | -0.05 (0.43) | -0.13 | -0.44 | -0.28 | 0.14 | 0.42 |
| Trial B short form | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| A | 0.92 (0.02) | 0.91 | 0.90 | 0.91 | 0.93 | 0.94 |
| b | 0.06 (0.29) | -0.10 | -0.11 | -0.11 | 0.15 | 0.40 |
| FE (N = 3) | | | | | | |
| Trial A short form | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| A | 0.96 (0.03) | 0.96 | 0.93 | 0.94 | 0.97 | 0.99 |
| b | -0.20 (0.05) | -0.22 | -0.24 | -0.23 | -0.18 | -0.15 |
| Trial B short form | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| A | 0.93 (0.05) | 0.91 | 0.89 | 0.90 | 0.94 | 0.98 |
| b | -0.29 (0.30) | -0.28 | -0.60 | -0.44 | -0.14 | 0.00 |

Differences between *A* and *b* scores by trial are shown for SMs in Figure 33 and Figure 34 and FEs in Figure 35 and Figure 36. All SMs and FEs performed above the average control sensitivity and all three groups demonstrated an overall decline in sensitivity on trial B short form. For response bias, both SMs and FEs showed a smaller range of *b* scores than controls, although individual scores of *b* differed between trials.

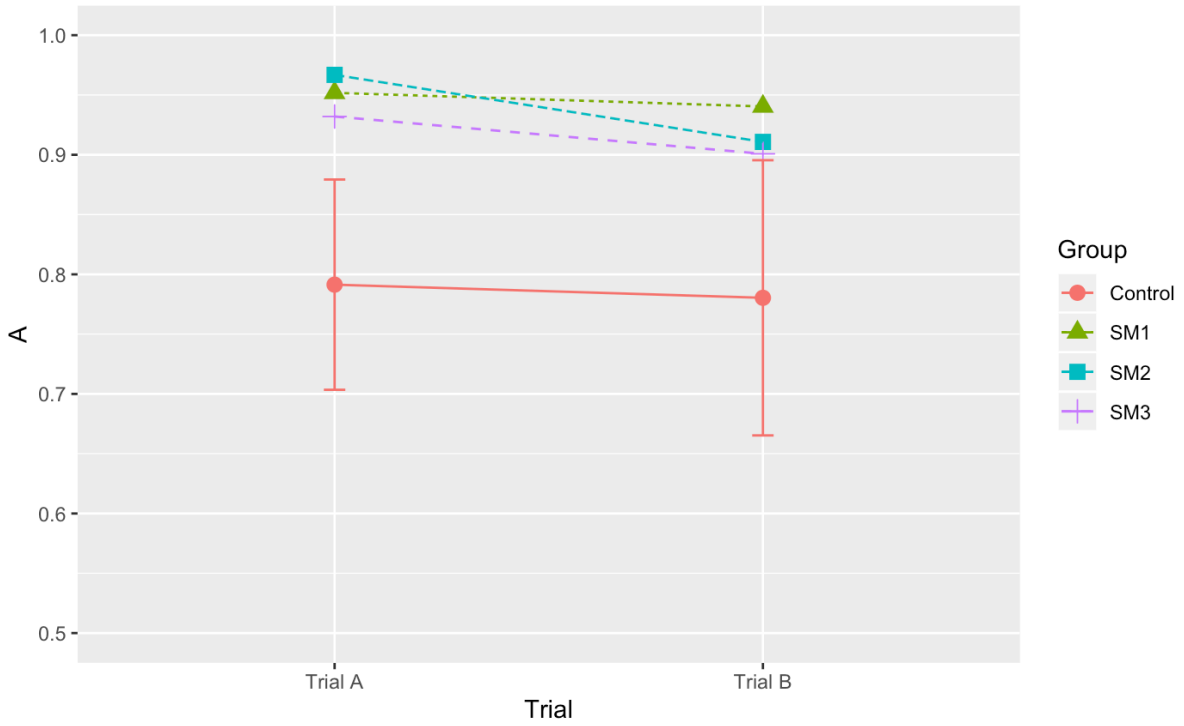


*Figure 33 – Sensitivity (A) by trial for individual SMs and controls (errors bars represent one standard deviation from the control mean)*

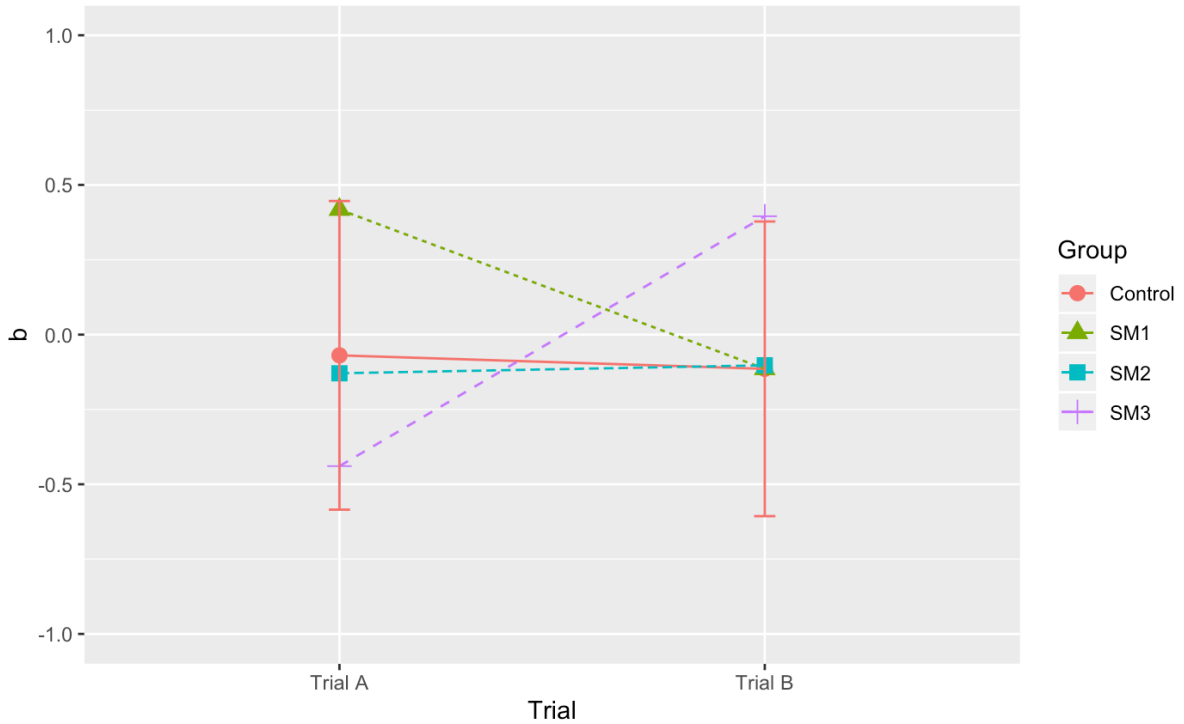


*Figure 34 – Response bias (b) by trial for individual SMs and controls (errors bars represent one standard deviation from the control mean)*

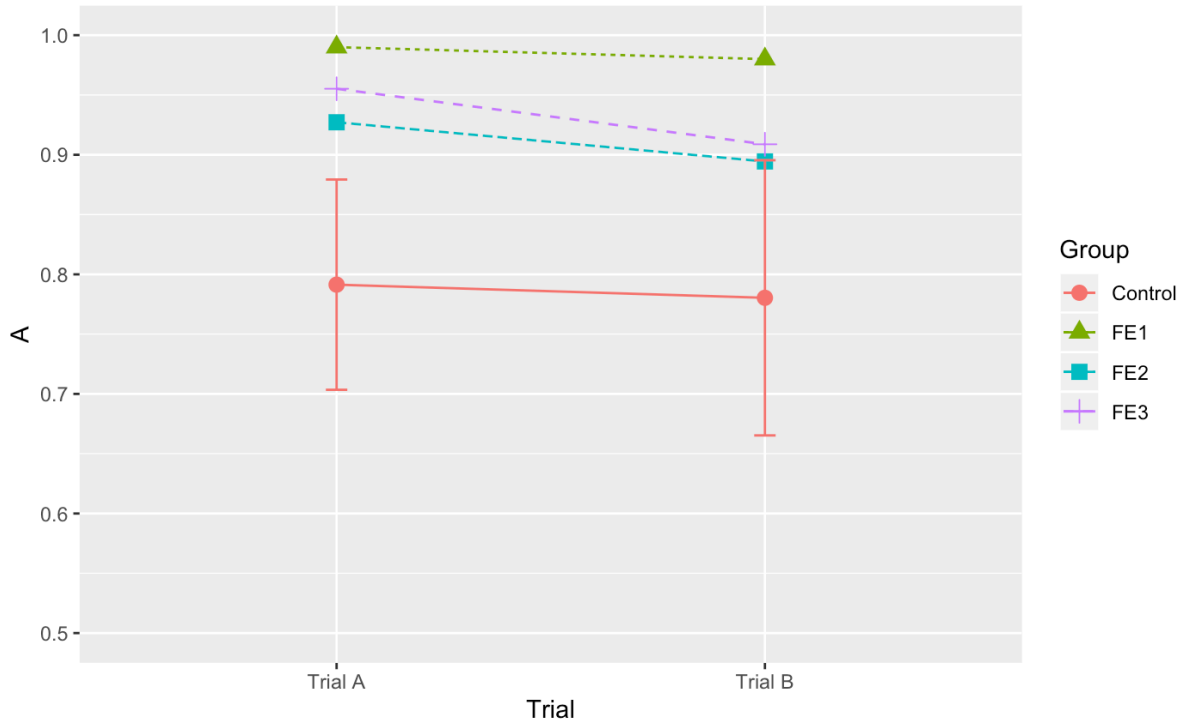


*Figure 35 – Sensitivity (A) by trial for individual FEs and controls (errors bars represent one standard deviation from the control mean)*

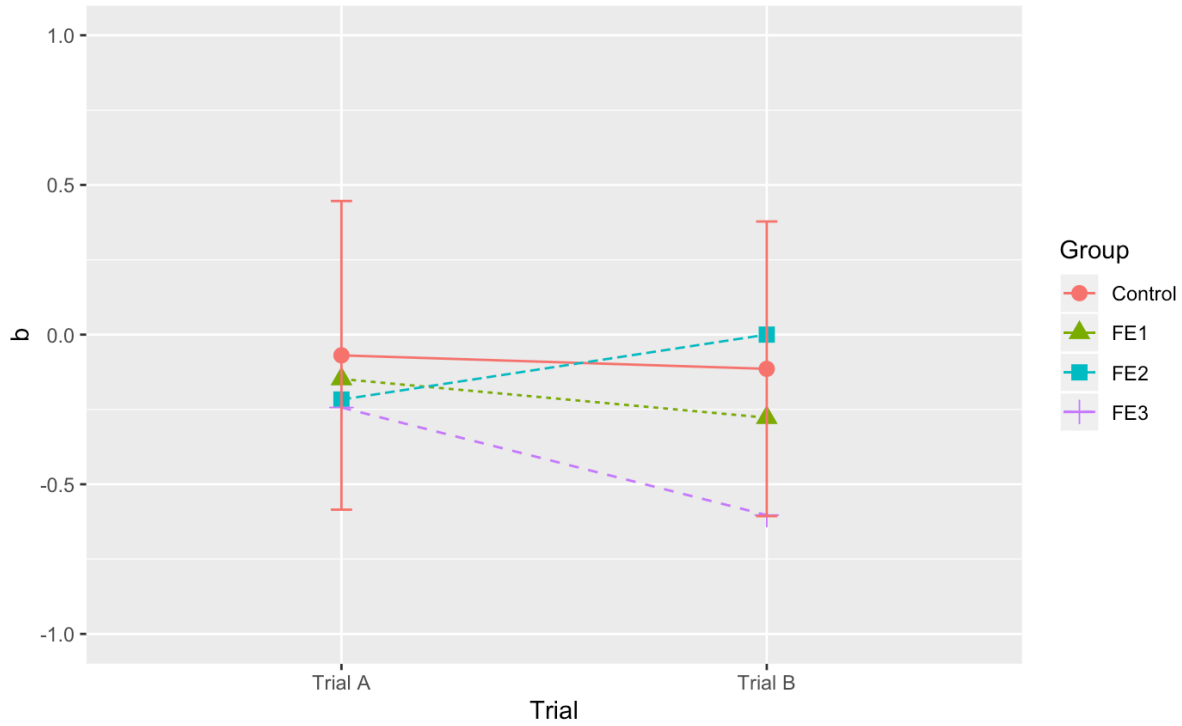


*Figure 36 – Response bias (b) by trial for individual FEs and controls (errors bars represent one standard deviation from the control mean)*

A factorial ANOVA of *A* scores revealed a significant effect for group ($F(2,28) = 8.11$, $\eta_p^2 = .305$, $p = .002$). No significant effect was found for trial ($F(1,28) = 0.87$, $\eta_p^2 = .007$, $p = .359$) and no significant interaction between trial and group ($F(2,28) = 0.11$, $\eta_p^2 = .002$, $p = .893$). Post hoc pairwise tests with Bonferroni adjustment for group revealed a similar pattern seen for overall accuracy, with a significant difference on trial A short form between FEs and controls ($t(44.2) = -2.85$, $d = 1.96$, $p = .020$) and SMs and controls ($t(44.2) = -2.72$, $d = 1.88$, $p = .028$). For trial B short form the difference between FEs and controls were significant ($t(44.2) = -2.52$, $d = 1.32$, $p = .046$), however no other group comparisons were statistically significant. A factorial ANOVA for *b* revealed no significant effect for group ($F(2,28) = 0.27$, $\eta_p^2 = .001$, $p = .762$) or trial ($F(1,28) = 0.00$, $\eta_p^2 = .001$, $p = .944$) and no interaction between trial and group ($F(2,28) = 0.19$, $\eta_p^2 = .002$, $p = .824$).

Individual difference analyses between trial A short form and trial B short form revealed that for measures of A, SMs (Table 38) and FEs (Table 40) were consistent between trials with no significant differences compared to controls. Individual SMs and FEs varied in response bias between trials (see Figure 34 and Figure 36). For all FEs these differences were within the variability of controls (Table 41). For SM1 and SM2 differences in b between trials were not statistically significant. For SM3 the difference in response bias was significant for the two-tailed test (Table 39). SM3 shifted from a negative value of *b* to a positive value, showing a change in face matching response bias from a greater proportion of match responses on trial A short form to a greater proportion of non-match responses on trial B short form. Interestingly, this shift in response did not significantly affect SM3's sensitivity between trials.

*Table 38 – Individual difference analyses comparing sensitivity (A) of superior face matchers between trials with mean control A*

| | Mean *A* (SD) | SM1 | SM2 | SM3 |
|---|---|---|---|---|
| *A* (trial A short form) | - | 0.95 | 0.97 | 0.93 |
| *A* (trial B short form) | - | 0.94 | 0.91 | 0.90 |
| | | | | |
| Control (N = 25) | | | | |
| Trial A short form | 0.79 (0.09) | | | |
| Trial B short form | 0.78 (0.12) | | | |
| *t* (24) | - | 0.46 | 0.95 | 0.57 |
| p (one-tailed) | - | .311 | .185 | .291 |
| p (two-tailed) | - | .662 | .370 | .581 |
| 95% CI | - | [12.51, 58.36] | [4.62, 40.06] | [11.52, 51.13] |
| Population with greater score discrepancy than individual (%) | - | 33.11 | 18.51 | 29.06 |

*Table 39 – Individual difference analyses comparing response bias (b) of superior face matchers between trials with mean control b*

| | Mean *b* (SD) | SM1 | SM2 | SM3 |
|---|---|---|---|---|
| *b* (trial A short form) | - | 0.42 | -0.13 | -0.44 |
| *b* (trial B short form) | - | -0.11 | -0.10 | 0.40 |
| | | | | |
| Control (N = 25) | | | | |
| Trial A short form | -0.07 (0.52) | | | |
| Trial B short form | -0.11 (0.49) | | | |
| *t* (24) | - | 1.13 | -0.15 | -2.08 |
| p (one-tailed) | - | .136 | .440 | .024 |
| p (two-tailed) | - | .271 | .880 | .048 |
| 95% CI | - | [4.68, 26.95] | [29.16, 59.46] | [0.24, 7.97] |
| Population with greater score discrepancy than individual (%) | - | 13.56 | 43.98 | 2.40 |

*Table 40 – Individual difference analyses comparing sensitivity (A) of face examiners between trials with mean control A*

| | Mean *A* (SD) | FE1 | FE2 | FE3 |
|---|---|---|---|---|
| *A* (trial A short form) | - | 0.99 | 0.93 | 0.96 |
| *A* (trial B short form) | - | 0.98 | 0.89 | 0.91 |
| | | | | |
| Control (N = 25) | | | | |
| Trial A short form | 0.79 (0.09) | | | |
| Trial B short form | 0.78 (0.12) | | | |
| *t* (24) | - | 0.57 | 0.95 | 0.83 |
| p (one-tailed) | - | .296 | .185 | .215 |
| p (two-tailed) | - | .592 | .370 | .429 |
| 95% CI | - | [8.28, 53.34] | [4.62, 40.06] | [6.18, 43.62] |
| Population with greater score discrepancy than individual (%) | - | 29.62 | 18.51 | 21.46 |

*Table 41 – Individual difference analyses comparing response bias (b) of face examiners between trials with mean control b*

| | Mean *b* (SD) | FE1 | FE2 | FE3 |
|---|---|---|---|---|
| *b* (trial A short form) | - | -0.15 | -0.22 | -0.24 |
| *b* (trial B short form) | - | -0.28 | 0.00 | -0.60 |
| | | | | |
| Control (N = 25) | | | | |
| Trial A short form | -0.07 (0.52) | | | |
| Trial B short form | -0.11 (0.49) | | | |
| *t* (24) | - | 0.23 | -0.61 | 0.80 |
| p (one-tailed) | - | .410 | .273 | .216 |
| p (two-tailed) | - | .820 | .545 | .432 |
| 95% CI | - | [26.14, 56.92] | [14.73, 42.36] | [9.20, 38.04] |
| Population with greater score discrepancy than individual (%) | - | 41.00 | 27.26 | 21.61 |

### 6.3.6. Confidence decisions of superior face matchers and face examiners

Previous studies have found that when forensic face examiners make errors, these decisions are made with low confidence (Norell et al., 2015; Phillips et al., 2018). The aim of analysing confidence ratings in this study was to understand whether the FEs in the current study would be similarly cautious in quick decision face matching compared to SMs and controls. As well as responding match or non-match to a face image pair, participants also had to rate their confidence in the decision using a four-point Likert scale ranging from 'not confident (1)' to 'extremely confident (4)'. Data were collapsed across trial A short form and trial B short form and the proportions of errors at each level of confidence were calculated for the three groups. The distributions of error proportions by confidence are shown in Figure 37. FEs made no extremely confident errors and only a very small proportion of very confident errors. Results for controls were more varied but showed a general trend for making a smaller proportion of extremely confident errors, this may be due to participants in these groups being recruited from a digital forensic department, making them perhaps more cautious than the non-police controls used in other studies. SMs showed large variations in confidence use and seldom made not confident errors, in contrast to FEs whose errors were predominantly not confident decisions.

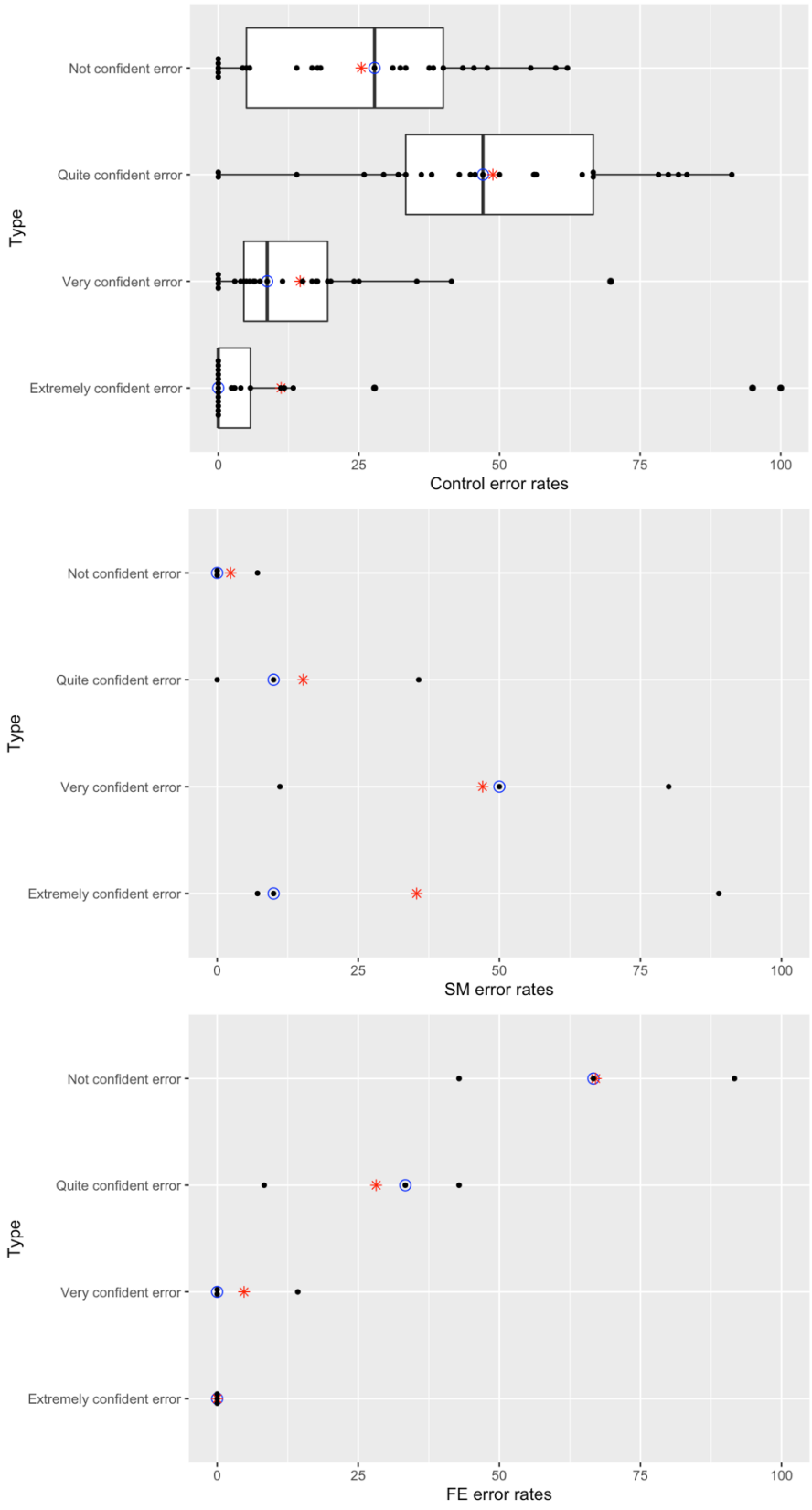


*Figure 37 – Boxplots of proportion of errors rate by confidence for each group*

Due to the non-normal distributions of error rates Kruskal-Wallis tests were used to compare differences in proportions of errors by confidence between the groups. No significant difference was found between the proportions of quite confident errors ($\chi^2(2, 30) = 5.78$, $p = .056$, $\varepsilon^2 = .18$) or very confident errors ($\chi^2(2, 28) = 5.34$, $p = .069$, $\varepsilon^2 = .17$). A significant difference was found between the proportions of not confident errors ($\chi^2(2, 28) = 8.74$, $p = .013$, $\varepsilon^2 = .28$). Post hoc tests using Dunn's test with Bonferroni correction revealed a significant difference between FEs and SMs ($p = .010$) but not between FEs and controls ($p = .079$) or SMs and controls ($p = .278$). A significant difference was also found between the proportions of extremely confident errors ($\chi^2(2, 28) = 6.05$, $p = .049$, $\varepsilon^2 = .20$). Post hoc tests using Dunn's test with Bonferroni correction revealed a significant difference between FEs and SMs ($p = .044$) but not between FEs and controls ($p = .475$) or SMs and controls ($p = .192$). These results demonstrate that FEs are more likely than SMs to make 'not confident' errors, whereas the SMs are more likely than FEs to make 'extremely confident' errors. Extreme confidence errors were particularly common for one of the SMs, highlight greater diversity in confidence decisions for this group compared to FEs.

### 6.3.7. Crowd effects

The final analysis in this Chapter investigated the effects of different sized crowds of individual SMs and FEs on accuracy and confidence, using trial B short form. Face-matching decisions were averaged between SMs and between FEs to create three face-matching pairs and one face-matching triad for each group. To prevent the occurrence of ambiguous responses encountered by White et al. (2013), where one individual in a pair responds match and the other non-match, decisions were converted to an eight-point Likert scale using the confidence ratings for each pair. Instead of a binary match/non-match decision, responses ranged from 'extremely confident non-match (1)' to 'extremely confident match (8)'. For match pairs, average responses of 5 (not confident match) or more

were deemed a correct response, whilst for non-match pairs average responses of 4 (not confident non-match) or less were deemed correct.

Table 42 shows summary statistics for the accuracy of SM and FE crowds on trial B short form. For SMs and FEs, pairs and triads showed an overall accuracy advantage. However, for FEs none of the crowds exceeded the accuracy of the top performing individual FE in a crowd. Crowd effects for SMs were more pronounced, with all pairs performing at the same level or greater than the accuracy of the top performing individual SM in the pair. The top performing SM pair had a slightly higher overall accuracy than the SM triad, shown by the distribution of accuracy scores in Figure 38.

*Table 42 – Summary statistics of accuracy for SM and FE crowds on trial B short form*

| Trial B short form | | | | | | |
|---|---|---|---|---|---|---|
| SMs | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Overall | 86.67 (3.06) | 86.00 | 84.00 | 85.00 | 88.00 | 90.00 |
| Matches | 85.33 (8.33) | 88.00 | 76.00 | 82.00 | 90.00 | 92.00 |
| Non matches | 88.00 (4.00) | 88.00 | 84.00 | 86.00 | 90.00 | 92.00 |
| SM Pairs | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Overall | 93.33 (2.31) | 92.00 | 92.00 | 92.00 | 94.00 | 96.00 |
| Match | 97.33 (2.31) | 96.00 | 96.00 | 96.00 | 98.00 | 100.00 |
| Non match | 89.33 (2.31) | 88.00 | 88.00 | 88.00 | 90.00 | 92.00 |
| SM Triad | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Overall | 94.00 | - | - | - | - | - |
| Matches | 96.00 | - | - | - | - | - |
| Non matches | 92.00 | - | - | - | - | - |
| FEs | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Overall | 88.00 (6.93) | 84.00 | 84.00 | 84.00 | 90.00 | 96.00 |
| Matches | 93.33 (8.33) | 96.00 | 84.00 | 90.00 | 98.00 | 100.00 |
| Non matches | 82.67 (10.07) | 84.00 | 72.00 | 78.00 | 88.00 | 92.00 |
| FE Pairs | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Overall | 91.33 (4.16) | 90.00 | 88.00 | 89.00 | 93.00 | 96.00 |
| Matches | 93.33 (6.11) | 92.00 | 88.00 | 90.00 | 96.00 | 100.00 |
| Non matches | 89.33 (4.62) | 92.00 | 84.00 | 88.00 | 92.00 | 92.00 |
| FE Triad | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Overall | 96.00 | - | - | - | - | - |
| Matches | 96.00 | - | - | - | - | - |
| Non matches | 96.00 | - | - | - | - | - |

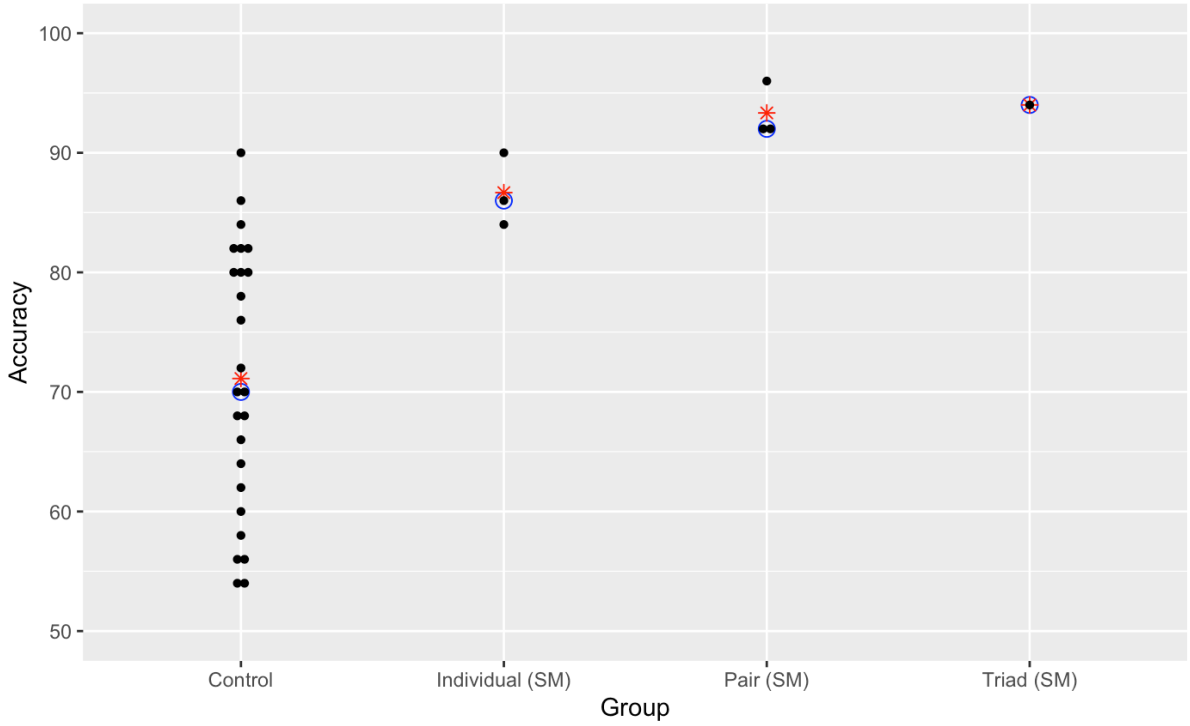


*Figure 38 – Accuracy on trial B short form for controls, SMs and SM crowds (blue circle represents median group score and red star is the mean group score)*

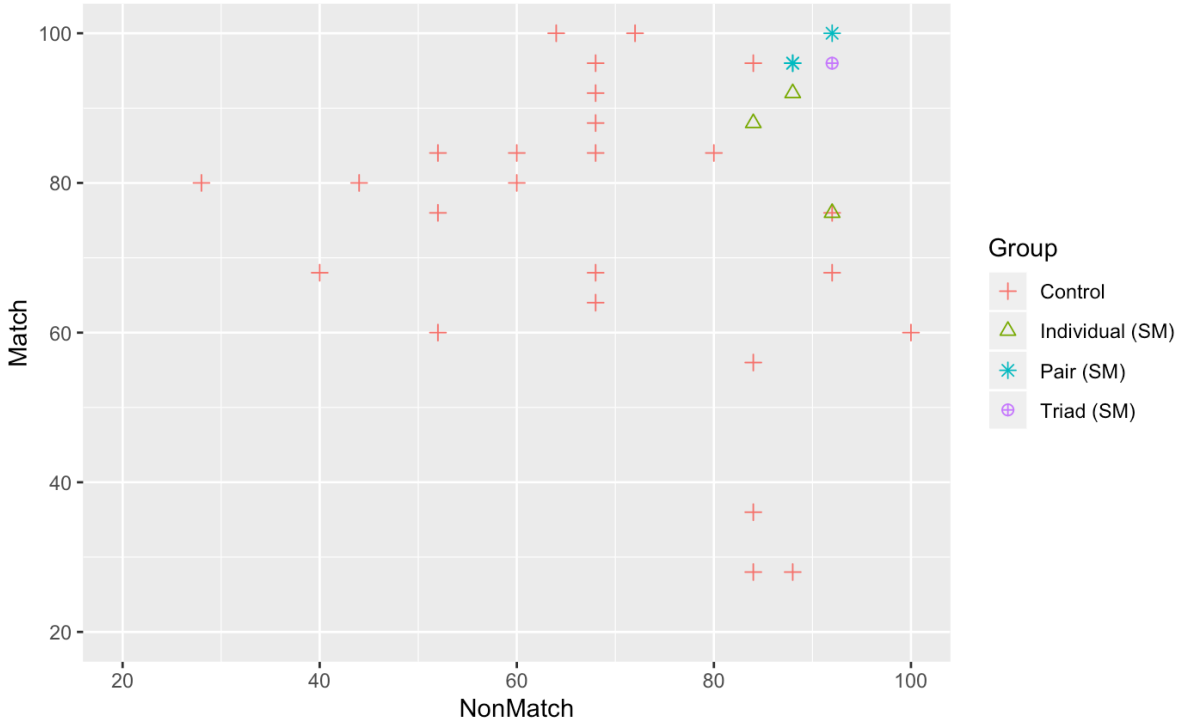


*Figure 39 – Scatterplot of match and non-match accuracy for controls, SMs and SM crowds on trial B short form*

Modified single-case t-tests were used to compare the overall accuracy of individual SM crowds to the mean accuracy of the control group on trial B short form (Table 43). All SM crowds outperformed controls at  a statically significant level using a one-tailed test. SM pair 3 was the top-performing crowd (p (two-tailed) = .037), however this pair did not include SM1, who was the top performing individual for this group. This suggests that for SMs, individual performance was not the sole driver of crowd effects and it may be the diversity of decisions between SM2 and SM3 that caused the large increase in overall accuracy seen for this pair.

*Table 43 – Individual case analyses comparing accuracy of SM crowds with mean control accuracy on trial B short form*

| | Mean overall accuracy (SD) | SM Pair 1 | SM Pair 2 | SM Pair 3 | SM Triad |
|---|---|---|---|---|---|
| Overall accuracy | - | 92 | 92 | 96 | 94 |
| Control (N = 25) | 71.12 (11.05) | | | | |
| *t* (24) | - | 1.89 | 1.89 | 2.25 | 2.07 |
| p (one-tailed) | - | .038 | .038 | .019 | .027 |
| p (two-tailed) | - | .076 | .076 | .037 | .054 |
| 95% CI | - | [88.89, 99.45] | [88.89, 99.45] | [93.30, 99.86] | [91.31, 99.72] |
| Population below individual's score (%) | - | 96.18 | 96.18 | 98.15 | 97.32 |

Figure 40 shows the distribution of accuracy scores for FE crowds compared to individual FEs and controls. This figure demonstrates that crowd effects for FEs are not as pronounced as for SMs. For one of the FE pairs, averaging responses resulted in an accuracy score that was less than the highest performing member of that pair. FE pair 1 is an average of responses from FE1 and FE2. FE1 achieved 96% accuracy on the task and

FE2 84% accuracy, however FE pair 1 had an overall accuracy of 90%, which is a decrease compared to the accuracy of FE1 individually.

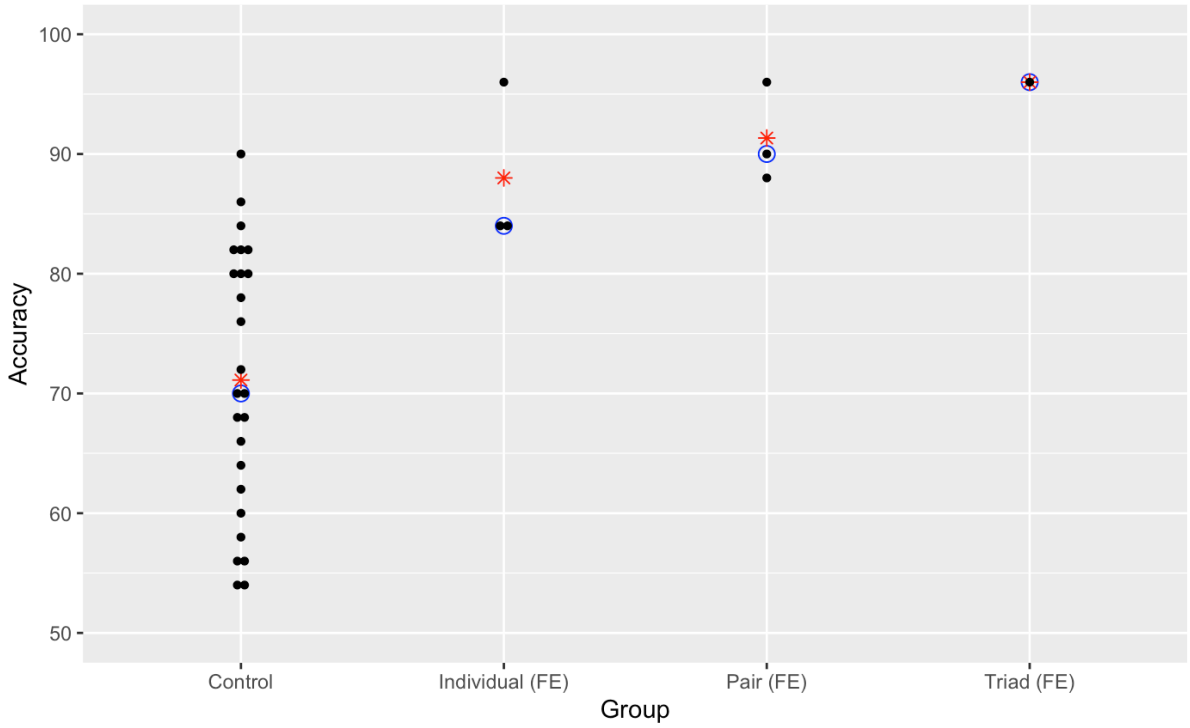


*Figure 40 – Accuracy on trial B short form for controls, FEs and FE crowds (blue circle represents median group score and red star is the mean group score)*

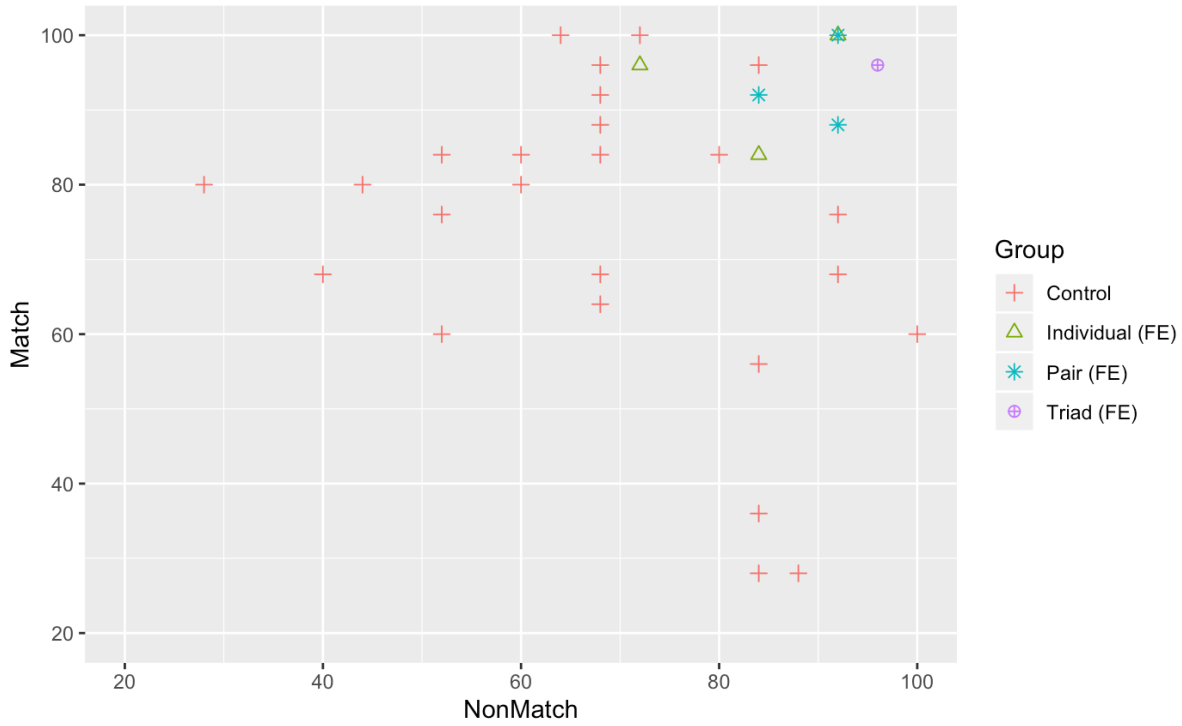


*Figure 41 – Scatterplot of match and non-match accuracy for controls, FEs and FE crowds on trial B short form*

Modified single-case t-tests revealed that only FE pair 2 and the FE triad outperformed controls at a statistically significant level. FE pair 1 bordered on significance for a one-tailed test, but as mentioned above, the accuracy of this pair was lower than the top performing individual member. It appears that crowd effects for FEs were being driven more by the almost ceiling levels of performance by FE1 rather than diversity in decision making, which appeared to be a bigger driver for the SM group.

*Table 44 – Individual case analyses comparing accuracy of FE crowds with mean control accuracy on trial B short form*

| | Mean overall accuracy (SD) | FE Pair 1 | FE Pair 2 | FE Pair 3 | FE Triad |
|---|---|---|---|---|---|
| Overall accuracy | - | 90 | 96 | 88 | 96 |
| Control (N = 25) | 71.12 (11.05) | | | | |
| *t* (24) | - | 1.71 | 2.25 | 1.53 | 2.25 |
| p (one-tailed) | - | .053 | .019 | .074 | .019 |
| p (two-tailed) | - | .107 | .037 | .147 | .037 |
| 95% CI | - | [86.00, 98.98] | [93.30, 99.86] | [82.58, 98.22] | [93.30, 99.86] |
| Population below individual's score (%) | - | 94.65 | 98.15 | 92.64 | 98.15 |

In the previous analysis individual SMs were observed to make significantly more extremely confident errors than FEs. Figure 42 displays the distribution of confidence decisions made by the three individual SMs, showing that most errors were either very confident or extremely confident. The distribution of confidence decisions for the three SM pairs are shown in Figure 43 to allow direct comparison to Figure 42. When SMs are combined into pairs there were no extremely confident errors and proportionally fewer errors were very confident.


*Figure 42 – Distribution of individual SM confidence decisions for trial B short form*

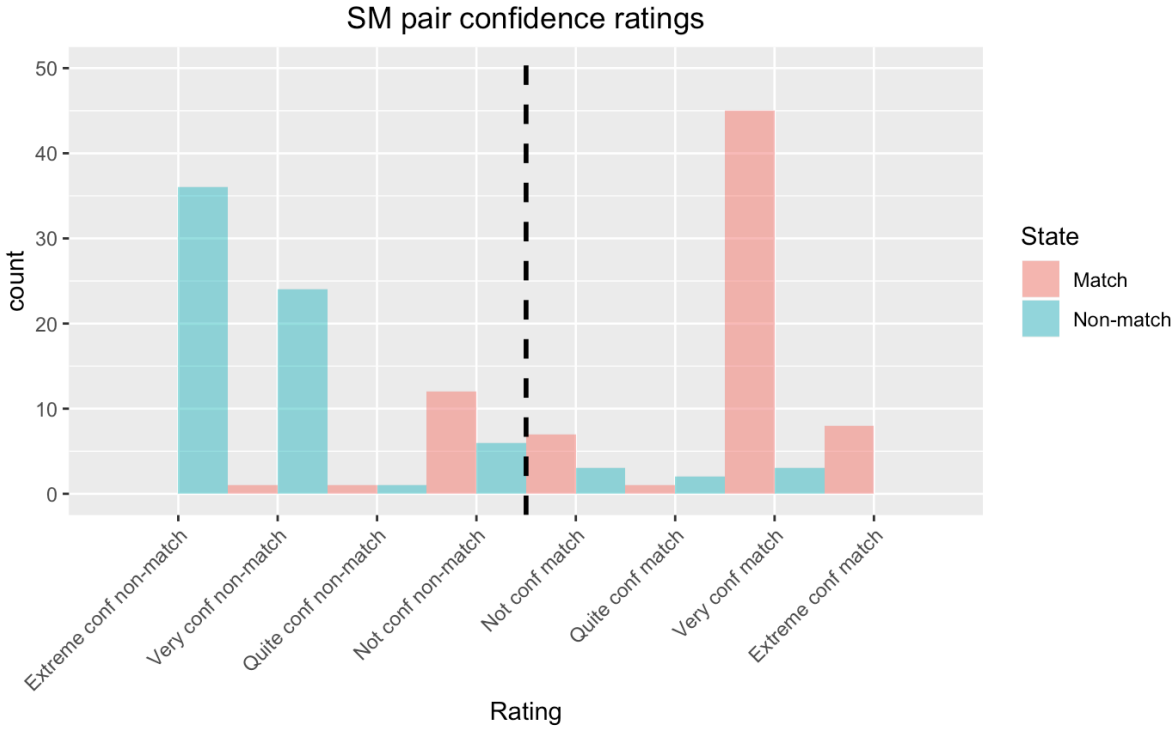


*Figure 43 – Distribution of SM pair confidence decisions for trial B short form*

## *6.4. Discussion*

The purpose of this study was to compare the perceptual skill of untrained superior face matchers and trained face examiners at quick decision face-matching tasks. Given that short face-matching training courses are largely ineffective at improving face-matching accuracy, trained examiners and individuals with high pre-existing face-matching ability may present a solution for improving face-matching performance in high throughput applied settings, such as at the border or in police investigations. Previous studies have demonstrated that face examiners have superior perceptual skills when quickly matching faces (White, Phillips, et al., 2015), as do some untrained individuals (Bate et al., 2018). However, examiners undergo lengthy training in feature-based face-matching procedures that can last several years, making them an expensive and scarce resource for face-matching agencies. Examiners are also not typically used in high-throughput face-matching settings like the border, and instead carry out detailed examinations taking several hours or even days (Moreton, 2021). If similar levels of performance in face-matching accuracy could be achieved simply by selecting high-performing superior face matchers, this would provide agencies with an efficient and effective way to improve face-matching accuracy in applied settings. In this study, three superior face matchers, selected from a pool of 28 candidates, achieved similar levels of face-matching accuracy at re-test as three trained face examiners. Where examiners did outperform the superior face matchers this was largely due to one high performing examiner, which demonstrates the importance of individual level analysis for future research of both superior face matchers and face examiners.

Researchers have suggested that there are qualitative differences in the perceptual face-matching skills of face examiners compared to novices (Towler et al., 2021; White, Phillips, et al., 2015), however, in this study face examiners were marked more by their similarities

to untrained superior face matchers rather than their differences. Individual examiners varied in accuracy and not all were statistically superior to controls. Examiners also varied in sensitivity and response bias across repeated tests, in a similar fashion to controls and superior face matchers. The study did find qualitative differences in the use of confidence ratings by examiners compared to superior face matchers. When examiners made errors, they were most commonly not confident in the decision and did not make any extremely confident errors. Whereas the superior face matchers seldom made not confident errors and made a significantly higher proportion of extremely confident errors, there were also large individual differences in confidence decisions between superior face matchers. This highlights a potential risk if agencies use superior face matchers in applied settings without additional safeguards to protect against errors being made with high levels of confidence, as such errors could have profound and life-changing consequences.

The three superior face matchers selected in this study did retain an accuracy advantage over controls at re-test but this was not statistically significant at the group or individual level. Therefore, selecting superior face matchers from small cohorts using a single test is unlikely to give significant gains in accuracy, replicating findings by Balsdon et al. (2018). In the current study the selection test and re-test were strongly correlated. In applied settings selection tests may not be as strongly correlated to real world face-matching tasks, which could further diminish the benefits of selection using a single test. Also, the superior face matchers were selected from a small pool of individuals. Selecting from larger pools of candidates would increase the probability of choosing exceptional individuals. However, for agencies where resources and personnel are limited this may not be possible. Using the wisdom of the crowd to combine decisions from the three superior face matchers resulted in face-matching pairs and a triad that were all statistically superior to controls. Also, by averaging responses the crowds of superior face matchers did not make any extremely high

confidence errors, demonstrating another benefit of the wisdom of the crowds approach. Interestingly, crowds were not as effective for face examiners in this study, which may be due to less diversity in the face matching decisions of examiners.

The results from this study demonstrate that combining face-matching selection tests, which are representative of the intended task, with a wisdom of the crowd approach can result in significant gains in face-matching accuracy. These gains were found when selecting superior face matchers from small groups of less than 30 individuals. These findings present a resource effective approach that not only improves accuracy on quick decision face-matching tasks, but also reduces the risk of high confidence errors being made.

# 7. Study Four – Combining human and algorithm face matching expertise

## 7.1. Introduction

The studies reported in previous chapters of this thesis have focussed on human face-matching accuracy and behaviour. However, automated facial recognition technology is now widely used in real world face-matching scenarios. For example, e-gates and biometrically enabled passports are commonplace for verifying traveller identity at the UK border and UK police forces are able to search custody image databases using automated technology. The accuracy of automated facial recognition algorithms has increased rapidly in recent years (Masi et al., 2019) but still requires supervision and monitoring by human operators (Stevens, 2021). Often, in applied face-matching systems, the result from an automated algorithm will be passed to a human operator for review, who then carries out a visual comparison to decide whether the images are a match (Towler, Kemp, et al., 2017). A study simulating this approach found that trained facial reviewers introduced significant errors when verifying face-matching results from an algorithm (White, Dunn, et al., 2015). As automated facial recognition technology continues to increase in accuracy, verification tasks for facial reviewers will also become increasingly more challenging (Academy of Social Sciences in Australia Inc., 2020). There is, therefore, a pressing need to research and design more effective ways of integrating automated technology and human operators in applied face-matching systems.

Laboratory studies have found that fusing results from facial recognition algorithms and humans on a face-matching task can introduce significant gains in accuracy (O'Toole et al.,

2007; Phillips et al., 2018). These gains are understood to be driven by the diverse face-matching strategies used by algorithms and humans (O'Toole et al., 2007), and are most pronounced when fusing state of the art algorithm scores with the face matching decisions of face examiners and super recognisers (Phillips et al., 2018). Human-algorithm fusion, therefore, appears to be an effective technique for improving accuracy in applied face-matching systems.

The aim of this study was to understand the effect of fusion at different levels of human and algorithm performance, using human participants of varying levels of face-matching ability and using face pairs that are challenging to human observers and face pairs that are challenging to an automated facial recognition algorithm.

## *7.2. Method*

### 7.2.1. Participants

The group of 138 police officers and staff from a UK police force (39 female) used as participants in Chapter 6 of this thesis were also used as participants in this study. A proprietary implementation of a DCNN facial recognition algorithm trained using the VGGFace2 dataset and developed by Qumodo Ltd. was also tested in this study. For details of the open source implementation of VGG Face2 dataset and models see Cao et al. (2018).

### 7.2.2. Materials

All 138 participants in the human group and the facial recognition algorithm matched images from trial A, consisting of 107 face pairs (54 matching pairs and 53 non-matching pairs), previously used in Chapters 5 and 6 of this thesis.

### 7.2.3. Procedure

Human participants completed face-matching trial A online using Qualtrics in the participants' place of work, using the same procedure for controls documented in Chapter 5. Prior to completing the trials participants consented to take part in the study. The same face pairs from trial A were processed by the facial recognition algorithm, resulting in a similarity score signifying how similar two faces are. Comparing two identical images would result in a score of zero, as the faces become more dissimilar the similarity score increases, therefore lower scores are more indicative of a match.

In order to allow the fusion of human face-matching decisions and algorithm similarity scores, human decisions were converted from a binary match/non-match response to an eight point Likert scale using the match and non-match response confidence ratings. This

resulted in a face-matching decision scale ranging from extremely confident non-match (1) to extremely confident match (8). Algorithm scores were converted to negative values to follow the same numerical direction as human decision scores, the inverted similarity scores were then scaled to the human scores following the procedure used by Phillips et al. (2018). The equation for the scaling is given below, where $\mu_{\mathrm{H}}$ and $\mu_{\mathrm{A}}$ are the mean and standard deviation of human decisions, $\sigma_{\mathrm{H}}$ and $\sigma_{\mathrm{A}}$ are the mean and standard deviation of the algorithm scores, $s_i$ is the similarity score for a given face image pair and $\widehat{s_i}$ is the scaled similarity score.

$$\widehat{s_i} = \sigma_{\mathrm{H}}\left(\frac{s_i - \mu_{\mathrm{A}}}{\sigma_{\mathrm{A}}} + \mu_{\mathrm{H}}\right)$$

Performance between human participants, the algorithm and fused participants was compared using the area under the receiver operating characteristic curve (AUC). AUC is a measure of the probability that a score or rating predicts the class of a stimulus, in this case whether face image pairs are a match or a non-match. It is a characteristic of the receiver operating characteristic (ROC) curve where an AUC value of 1 means that ratings are a perfect prediction of face-pair class (match or non-match) and values of .5 means scores predict matches and non-matches at the level of chance. AUC allows the comparison of human decision ratings and algorithm similarity scores without having to impose thresholds on algorithm scores that denote a match or non-match. AUC scores were calculated in R using the pROC package (Robin et al., 2011).

The study received a favourable opinion by the ethical committee of the Open University.

## *7.3. Results*

### 7.3.1. Fusion for human-challenging faces

The first part of the analysis tested the performance of the algorithm and the effects of fusion on face pairs that were challenging to humans. AUC scores for human participants and the algorithm were calculated for faces pairs from trial A short form, used in Chapter 6 of this thesis, which contains the 25 hardest match pairs and 25 hardest non-match pairs for human observers from trial A. Table 45 shows summary statistics for AUC by group for the human challenging face pairs.

*Table 45 – Summary statistics of AUC scores for the algorithm, humans and fusion results on human challenging images from trial A*

| Human challenging images | | | | | | |
|---|---|---|---|---|---|---|
| AUC | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Algorithm | .973 | - | - | - | - | - |
| Humans | .790 (.107) | .814 | .506 | .714 | .862 | .974 |
| Fusion | .955 (.037) | .962 | .789 | .938 | .981 | 1 |

The algorithm performed exceptionally well, with an AUC score comparable to that of the top performing human and far exceeding the average human score. The human participants showed a wide range in AUC scores, from almost chance to close to ceiling. Fusion of human decision ratings with the standardised algorithm scores resulted in large gains in performance. The distributions of human and fused AUC scores are shown in Figure 44, revealing that a large number of fused scores exceeded the performance of the algorithm, whereas without fusion only one human participant performed at the level of the algorithm. Due to the non-normal distribution of human AUC scores ($W = 0.86$, $p < .001$) a non-

parametric paired Wilcoxon signed-rank test was used to compare the effects of score fusion. The difference between human AUC scores and fused AUC scores was found to be statistically significant with a large effect size ($V = 49$, $p < .001$, $r = .86$).

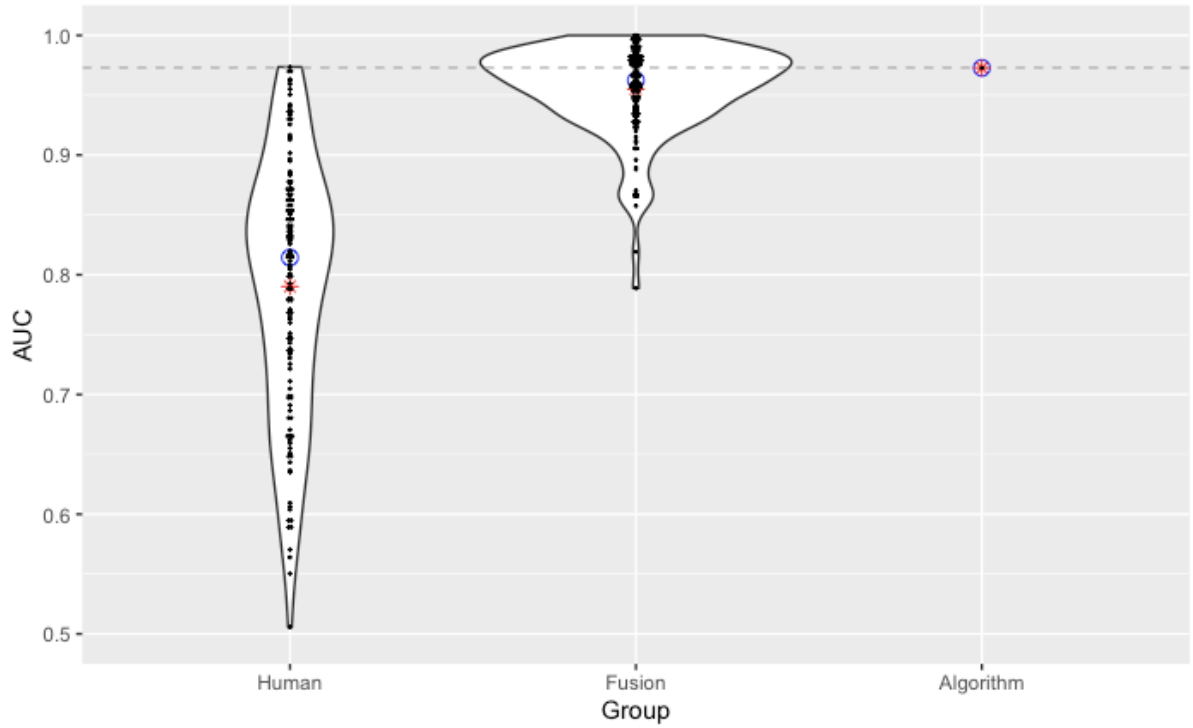


*Figure 44 – AUC scores by group for human challenging face pairs (grey dashed line represents algorithm score)*

Human AUC scores and fused AUC scores were strongly correlated, $r(136) = .790$, $p < .001$ (see Figure 45), demonstrating a positive linear relationship between human performance and the effects of fusion.

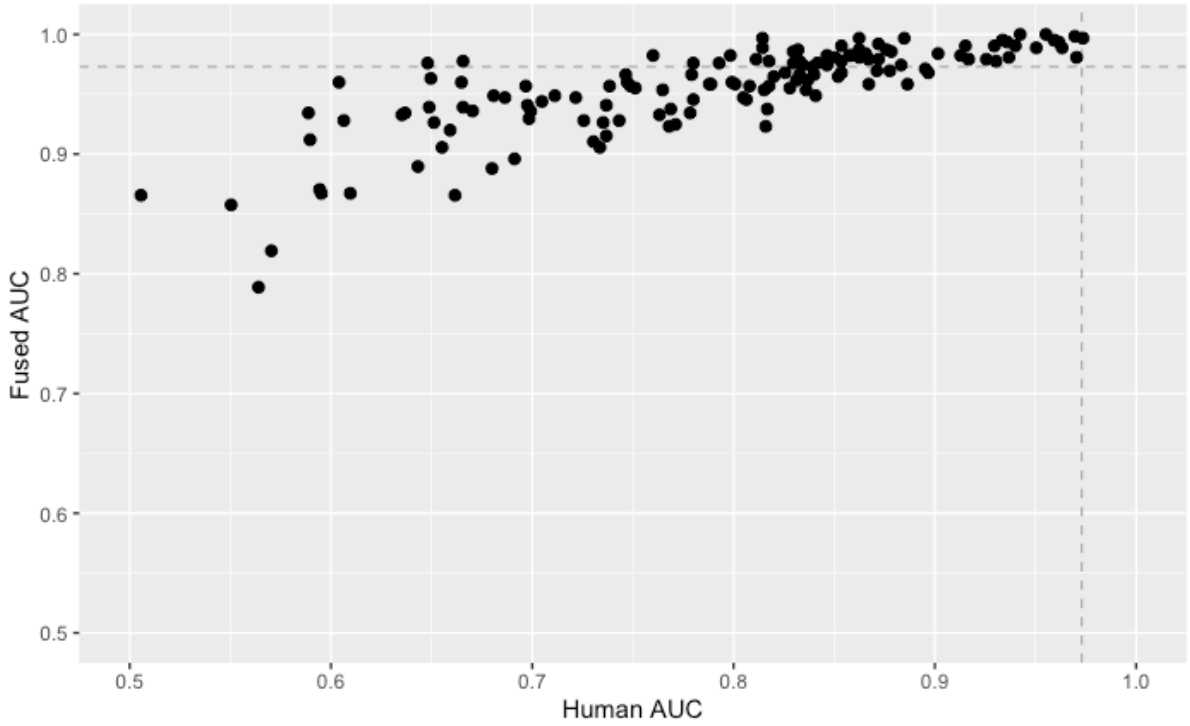


*Figure 45 – Scatterplot of human AUC scores and fused AUC scores for human challenging images (grey dashed lines represent algorithm score)*

To better the understand the relationship between fusion effects and human performance the human group was separated into three new groups consisting of the top 20 AUC scores, bottom 20 AUC scores and middle 20 AUC scores around the median, prior to fusion. Summary statistics for the three new human groups with and without fusion are shown in Table 46.

*Table 46 – Summary statistics of human and fused AUC scores split by performance on human challenging images from trial A*

| Human challenging images | | | | | | |
|---|---|---|---|---|---|---|
| Top 20 humans | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Unfused | .941 (.021) | .938 | .902 | .929 | .960 | .974 |
| Fused | .989 (.007) | .990 | .978 | .982 | .995 | 1 |
| Middle 20 humans | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Unfused | .811 (.011) | .814 | .789 | .803 | .818 | .830 |
| Fused | .963 (.018) | .959 | .923 | .955 | .976 | .997 |
| Bottom 20 humans | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Unfused | .610 (.041) | .608 | .506 | .589 | .648 | .659 |
| Fused | .903 (.048) | .916 | .788 | .867 | .934 | .976 |

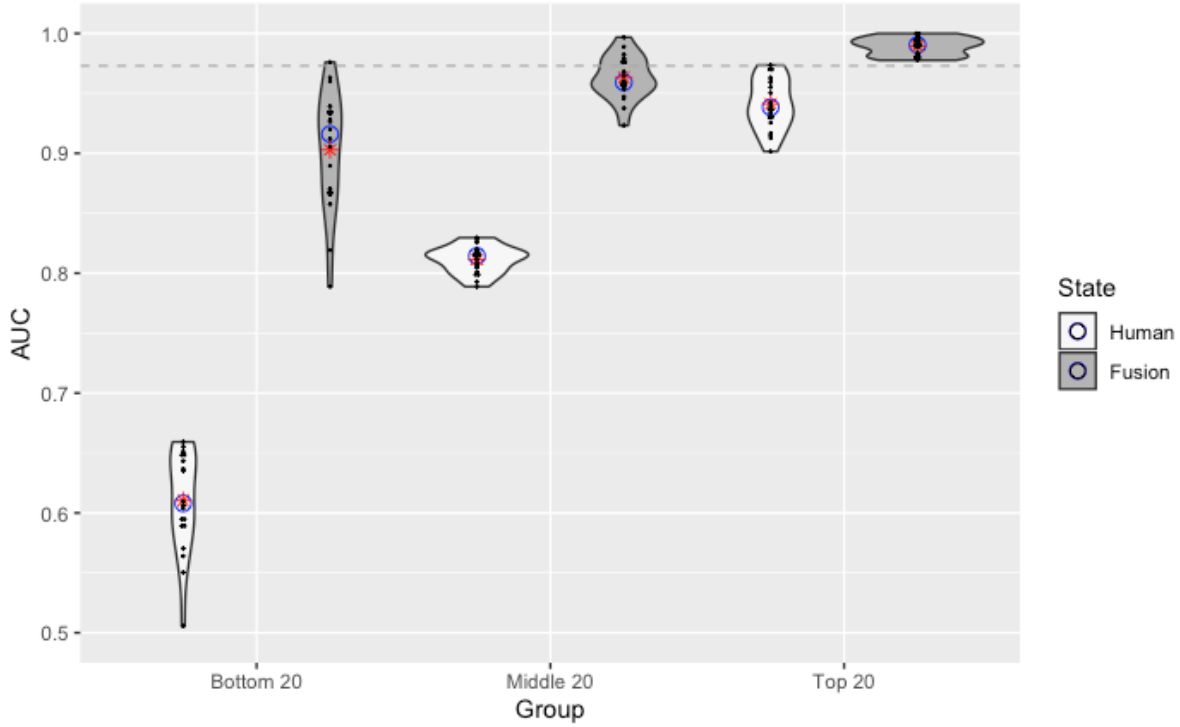


*Figure 46 – AUC scores for the human group separated by performance on human challenging face pairs (grey dashed line represents algorithm score)*

The distributions of unfused and fused AUC scores reveal that fusion caused large increases in performance for bottom, middle and top performers (Figure 46). All top performers exceeded the AUC score of the algorithm after fusion, whereas six middle performers and only one bottom performer exceeded the algorithm after fusion. Multiple Wilcoxon signed-rank tests with Bonferroni adjusted p-values revealed that the improvements from fusion were statistically significant for all groups ($V = 0$, $p < .001$, $r = -.88$ for all groups). A Kruskal-Wallis compared the fused AUC scores of the three groups, revealing a statistically significant difference with a large effect size ($\chi^2(2, 40) = 43.14$, $p < .001$, $\varepsilon^2 = .73$). Post hoc tests using Dunn's test with Bonferroni adjusted p-values revealed statistically significant differences between top and middle performers ($p = .002$), middle and bottom performers ($p = .005$) and top and bottom performers ($p < .001$) after fusion. The results demonstrate that the fusion of human ratings and algorithm similarity scores provides large gains in performance for face pairs that are challenging to human observers. Prior to fusion only one participant outperformed the algorithm, whereas after fusion 55 participants, almost 40% of the sample, outperformed the algorithm. Improvements from fusion had a linear relationship with human ability, meaning that highest post-fusion AUC scores were achieved with the human participants that were more accurate at face-matching to begin with, however gains were significant for top, middle and bottom performers.

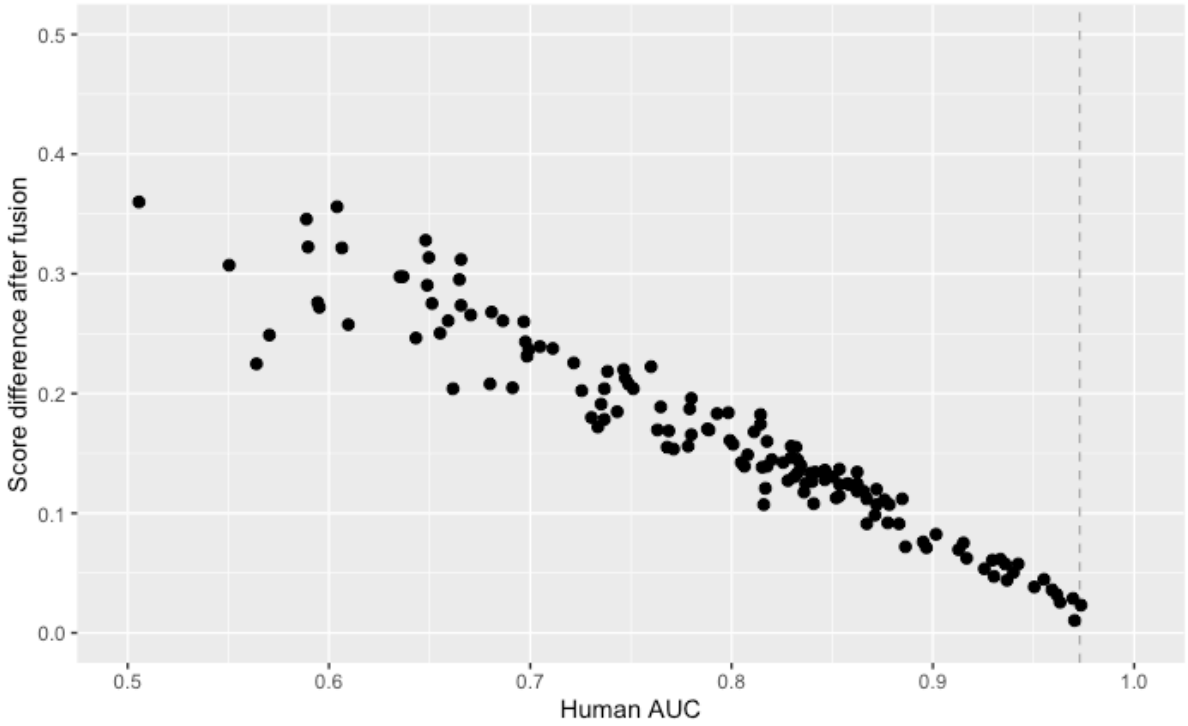


*Figure 47 – Scatterplot of human AUC scores and scores difference after fusion for human challenging images (grey dashed line represent algorithm score)*

AUC scores prior to fusion and the difference in score after fusion had a strong negative relationship ($r(136) = -.96$ $p < .001$) for face pairs that were challenging for humans. All human participants improved after fusion but the increase was largest for lower performing participants (Figure 47). As human performance increased, fusion caused more AUC scores to approach ceiling (Figure 45), thus limiting the extent of possible gains for the highest performing humans in this study.

### 7.3.2. Fusion for algorithm-challenging faces

The next stage of the analysis investigated the effects of fusion for face pairs that were challenging to the algorithm. Figure 48 shows the distribution of algorithm similarity scores for all matching and non-matching face pairs from trial A. Although the algorithm scores show good separation for the matching and non-matching pairs there was some overlap

between the two distributions. Face pairs within this overlapping region were selected as pairs that were challenging to the algorithm, resulting in 16 matching face pairs and 14 non-matching face pairs. The human decision ratings and algorithm similarity scores for these 30 face pairs were then fused.

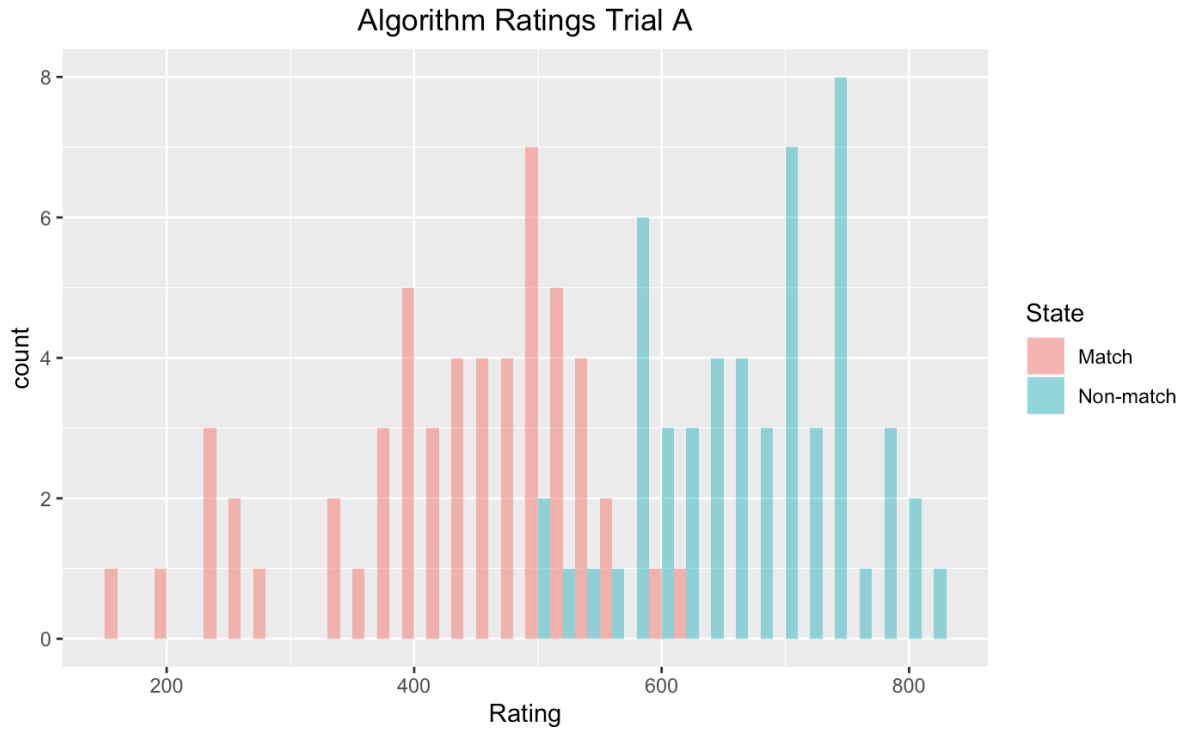


*Figure 48 – Frequency distribution of algorithm scores for matching and non-matching pairs on trial A*

*Table 47 – Summary statistics of AUC scores for the algorithm, humans and fusion on algorithm-challenging images from trial A*

| Algorithm challenging images | | | | | | |
|---|---|---|---|---|---|---|
| AUC | | | | | | |
| | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Algorithm | .710 | - | - | - | - | - |
| Humans | .850 (.097) | .863 | .549 | .790 | .921 | 1 |
| Fusion | .860 (.056) | .871 | .670 | .827 | .897 | .991 |

The algorithm performed poorly on these images, with an AUC score below the first quartile of the human AUC scores. The human participants again showed a wide range in performance, with one individual close to chance and another at ceiling. However, median and third quantile scores were higher than for the human-challenging faces, meaning the human participants were more accurate at matching face pairs that the algorithm struggled with. This suggests that the algorithm was matching face pairs in a qualitatively different way to humans.

Figure 49 shows the distributions of pre-fusion and fused AUC scores. It appears that, in this analysis, fusion affected the tails of the score distribution in different ways. The minimum and first quartile values increased after fusion whereas the third quartile and maximum values decreased. Due to the non-normal distribution of the human AUC scores ($W = 0.97$, $p = .003$) nonparametric tests were used to evaluate the effect of fusion on performance. A Wilcoxon signed-rank test revealed no statistically significant difference between pre-fusion and fused AUC scores ($V = 4457$, $p = .563$). However, all bar two of the fused AUC scores surpassed the algorithm when working alone, indicating that when the algorithm performs poorly, fusion with human face-matching decisions is more effective than the algorithm working independently.

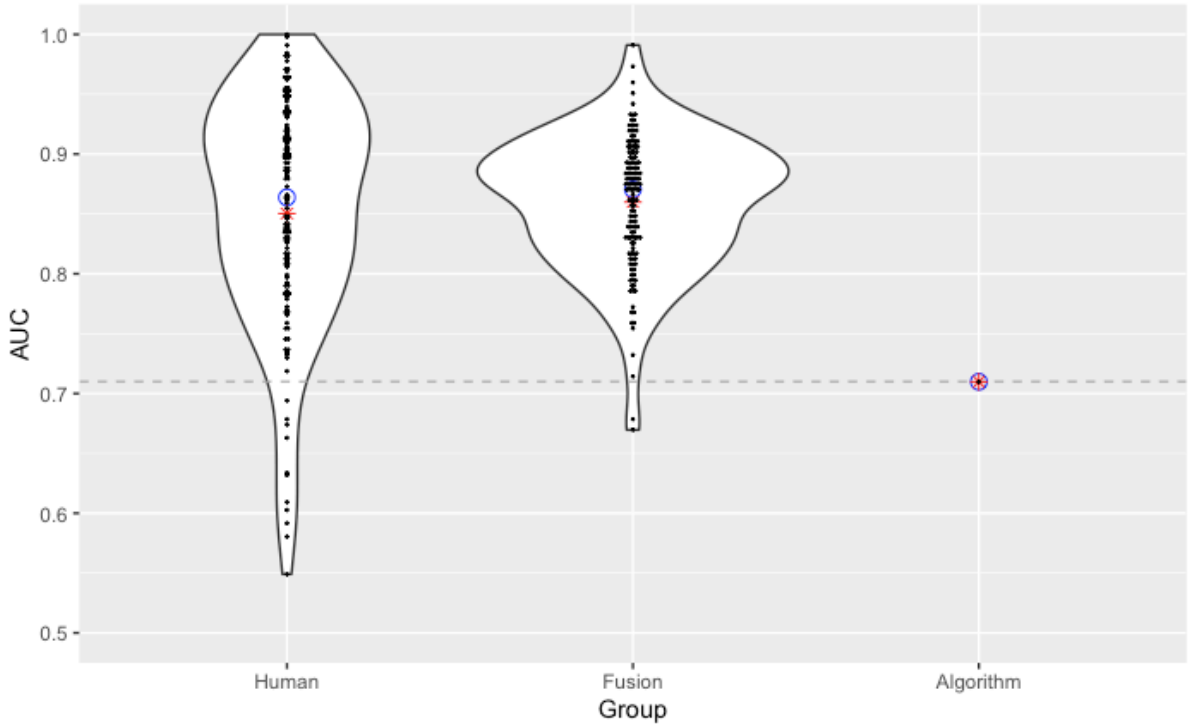


*Figure 49 – AUC scores by group for algorithm challenging face pairs (grey dashed line represents algorithm score)*

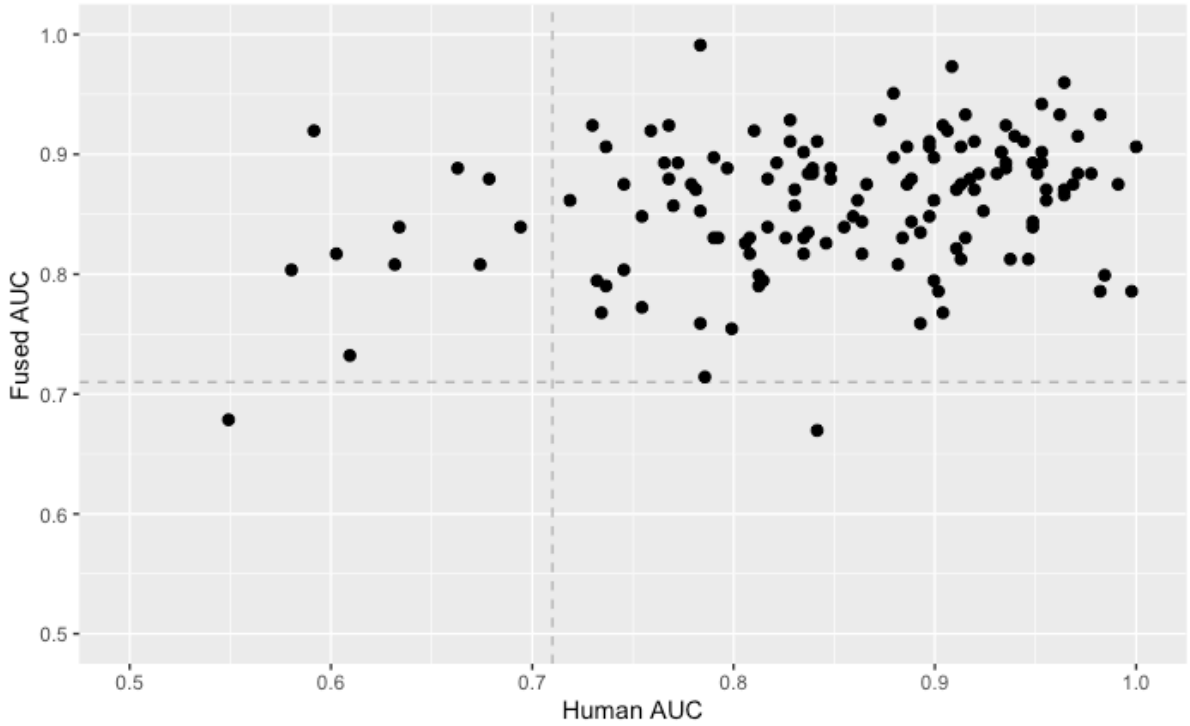


*Figure 50 – Scatterplot of human AUC scores and fused AUC scores for algorithm challenging images (grey dashed lines represent algorithm score)*

Pre-fusion and fused AUC scores for algorithm challenging face pairs had a weak to moderate positive relationship ($r(136) = .304$, $p < .001$) in contrast to the strong positive relationship observed for human challenging face pairs. It, therefore, appears that the effects of fusion are less consistent when the algorithm performs poorly, with some human participant AUC scores increasing after fusion and others decreasing (Figure 50).

The human group was split into the top 20 performers, middle 20 performers and bottom 20 performers using pre-fusion AUC scores for the algorithm-challenging face pairs. Summary statistics are shown in Table 48 and the distributions of AUC score pre- and post-fusion are shown in Figure 51. When the algorithm performed poorly the effects of fusion varied depending on the face-matching ability of the human participant. The AUC scores of the bottom 20 performers showed an overall improvement after fusion, which was statistically significant with a large effect size ($V = 0$, $p < .001$, $r = .88$). Fusion appeared to increase the range of AUC scores for middle performers with no overall improvement at the group level ($V = 44.5$, $p = .133$). For top performers there was a statistically significant decrease in AUC scores after fusion ($V = 0$, $p < .001$, $r = -.88$). The differing effects of fusion caused the distributions of AUC scores between the three groups to overlap after fusing, however, the differences between the fused groups were statistically significant ($\chi^2(2, 40) = 46.35$, $p < .001$, $\varepsilon^2 = .79$). Differences between the fused groups were confirmed with post hoc pairwise tests (middle 20 and bottom 20 $p < .001$, top 20 and middle 20, $p = .012$, middle 20 and bottom 20 $p < .001$, top 20 and bottom 20 $p < .001$).

*Table 48 – Summary statistics of human and fused AUC scores split by performance on algorithm challenging images from trial A*

| Algorithm challenging images | | | | | | |
|---|---|---|---|---|---|---|
| Top 20 humans | | | | | | |
| | Min | 1st Quartile | Median | Mean (SD) | 3rd Quartile | Max |
| Unfused | .951 | .955 | .967 | .970 (.015) | .982 | 1 |
| Fused | .871 | .896 | .924 | .922 (.032) | .935 | .911 |
| Middle 20 humans | | | | | | |
| | Min | 1st Quartile | Median | Mean (SD) | 3rd Quartile | Max |
| Unfused | .839 | .848 | .864 | .865 (.017) | .880 | .888 |
| Fused | .817 | .856 | .860 | .873 (.024) | .893 | .906 |
| Bottom 20 humans | | | | | | |
| | Min | 1st Quartile | Median | Mean (SD) | 3rd Quartile | Max |
| Unfused | .549 | .626 | .686 | .677 (.065) | .735 | .755 |
| Fused | .670 | .758 | .779 | .771 (.046) | .799 | .866 |

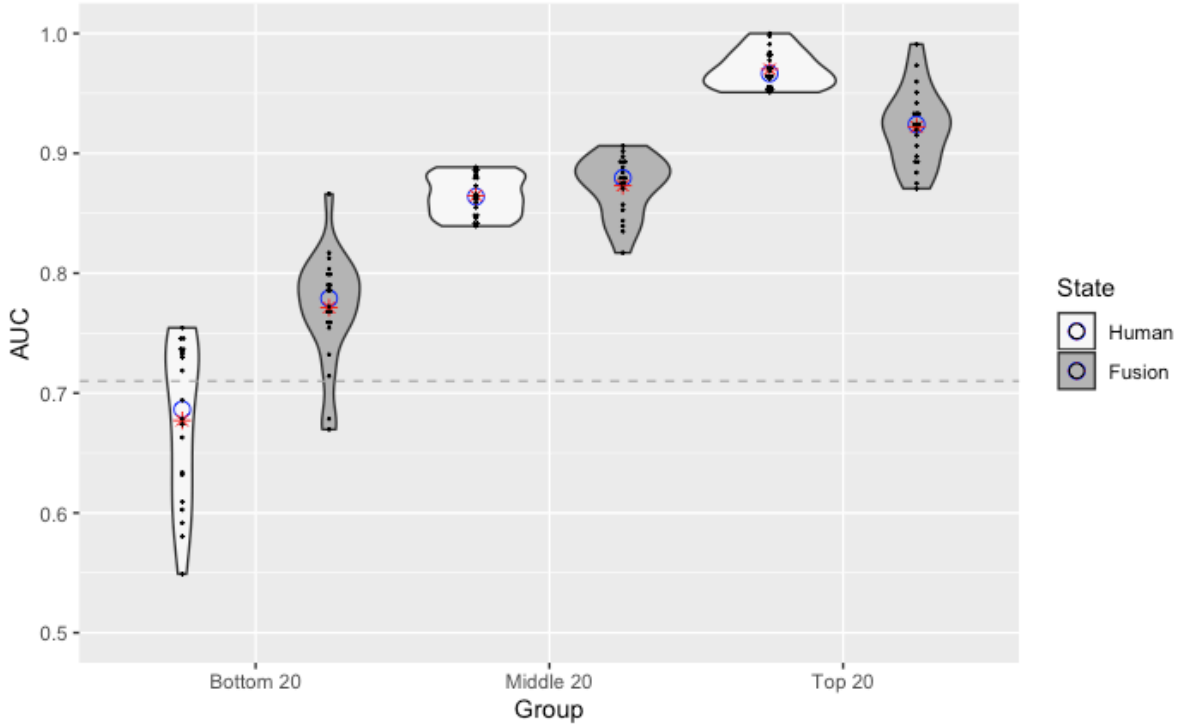


*Figure 51 – AUC scores separated by performance on algorithm challenging face pairs (grey dashed line represents algorithm score)*

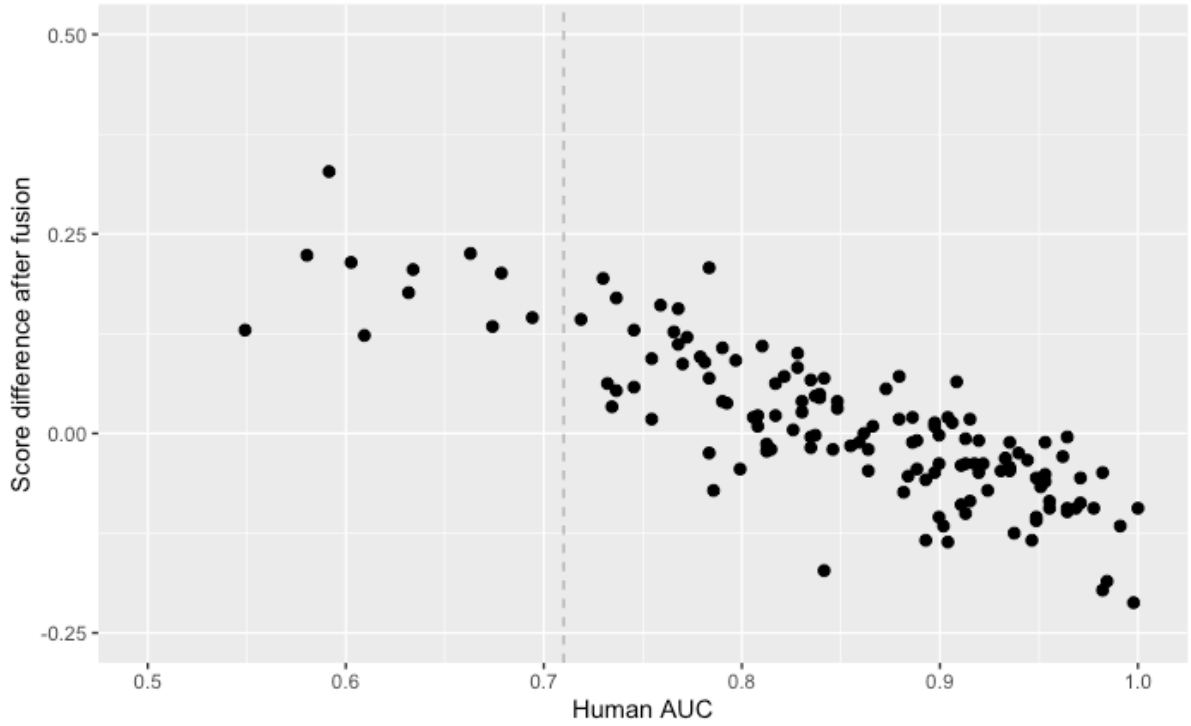


*Figure 52 – Scatterplot of human AUC scores and scores difference after fusion for algorithm challenging images (grey dashed line represent algorithm score)*

Human AUC scores prior to fusion and the difference in score after fusion had a strong negative relationship ($r(136) = -.83$ $p < .001$) for algorithm-challenging face pairs. Unlike the results for human-challenging face pairs, not all human participants improved after fusion (Figure 52). As human performance increased the likelihood that fusion would cause a decrease in performance was greater. This indicates that the accuracy of the algorithm and the accuracy of the human are major contributors to the benefits of fusion.

## *7.4. Discussion*

This study demonstrated that fusing human decision ratings and algorithm similarity scores can lead to significant improvements in face-matching performance, replicating findings from previous studies (O'Toole et al., 2007; Phillips et al., 2018). An overall benefit form fusion was observed for both human-challenging and algorithm-challenging face pairs. This suggests that independently combining human and machine face-matching decisions, rather than having the human operator adjudicate the algorithm results in a sequential fashion, could be highly effective for improving performance in applied settings. However, the effectiveness of the fusion was impacted by both human and algorithm performance.

Algorithm performance appeared to be a major driver of the benefits of fusion. For human challenging face pairs, where the algorithm performed very well, fusion resulted in gains in performance for all participants, however, for face pairs where the algorithm performed poorly the effects of fusion were more variable. Where the algorithm was less accurate, fusion increased AUC scores for low performers who were close to or below the algorithm score prior to fusion. For participants around the median AUC score fusion had mixed effects and for top performers fusion resulted in a decrease in AUC scores. It is important to note that these results were observed with the same algorithm and same human observers but on different face pairs. Therefore, the benefits of fusion are, to some extent, face-pair specific, providing gains for some face pairs but not others. This warrants caution in the wholesale use of human-machine fusion and indicates there may need to be additional safeguards or processes in place if fusion were to be used in applied settings.

The previous study in Chapter 6 of this thesis demonstrated that by crowd sourcing face-matching decisions from high performing individuals, statistically superior levels of accuracy could be achieved even when both individuals in the pair were not statistically superior

themselves. As a result, by combining face-matching selection tests with crowd sourcing techniques, less stringent criteria could be used select face-matching operators and superior performance still achieved. In the current study, results indicate that the effectiveness of fusion with a high performing algorithm could allow even less stringent selection criteria for operator selection, as some human participants performing around the median AUC score prior to fusion then surpassed the algorithm in performance after fusion. This would be beneficial for face-matching agencies where personnel numbers are too limited to allow selection of truly superior face-matchers using stringent cut offs, i.e. at two standard deviations above the population mean as suggested by Bobak, Pampoulov, et al. (2016).

Although algorithm performance appeared to be a major driver of fusion effects, there was some evidence for diversity in face-matching strategies between the human participants and the algorithm contributing to the fusion process. Firstly, human and algorithm performance varied on different face pairs. The algorithm performed exceptionally well for face pairs that the majority of human participants found challenging and human performance was greater for face pairs that challenged the algorithm. Secondly, for human-challenging face pairs improvements were observed even when algorithm scores were fused with ratings from humans who performed as well as or above the algorithm prior to fusion. This suggests that it was not solely the high performance of the algorithm that contributed to gains in accuracy from human-algorithm fusion, but also diversity in face matching strategies between the algorithm and humans, as suggested by O'Toole et al. (2007).

Both the Facial Identification Scientific Working Group and The National Institute for Standards and Technology recommend that human operators should review and adjudicate algorithm matches when face images are uncontrolled or low quality (FISWG, 2020; Grother

et al., 2019a). However, White, Dunn, et al. (2015) found that trained facial reviewers made significant numbers of errors when adjudicating algorithm results. More recently, Howard et al. (2020) demonstrated that algorithm face-matching judgements shifted the response bias of human face-matching decisions when observers were shown the algorithm decision. The results of these two studies suggest that having human operators review algorithm results in the sequential fashion suggested, as by FISWG and NIST, can introduce error and potentially bias the decisions of the human operator.

In the current study, human decision ratings and algorithm similarity scores were fused independently of each other, mitigating the risks of cognitive bias observed by Howard et al. (2020). Fusion also resulted in overall gains in performance by taking advantage of the diverse face-matching strategies used by humans and the algorithm, and this was most effective when the algorithm performed well. Despite being first demonstrated over 12 years ago by O'Toole et al. (2007) human-machine fusion does not appear to be widely adopted or understood within the applied face-matching community. For example, the Michigan State Police (MSP) Department quote the fusion results of Phillips et al. (2018) in an online FAQ[11] about their procedures for reviewing results from an automated facial recognition system. However, it appears that MSP procedure does not involve fusion and is instead based on the sequential review of the algorithm results by a trained examiner, which can be prone to bias and does not capitalise on the benefits of independent fusion. As the use of face recognition algorithms by police and other agencies increases and the task of human review becomes more challenging (Academy of Social Sciences in Australia Inc., 2020), researchers and practitioners must come together to design innovative and effective

[11] https://www.michigan.gov/documents/msp/Facial_Recognition_FAQ_666807_7.pdf

solutions for combining human and machine face-matching expertise. Based on the results of the current study fusion techniques look like a promising place to start.

# 8. Study Five – Operational accuracy of face examiners

## 8.1. Introduction

When testing the performance of forensic face examiners within their domain of expertise Towler et al. (2018) define two types of studies, those that test perceptual skill and those that test operational accuracy. Perceptual skill refers to the forensic practitioners raw ability in classifying or matching stimuli, without access to their standard procedures and tools. Operational accuracy refers to the performance of forensic practitioners in tests that replicate operational casework. Typically, in operational accuracy tests, the practitioners have access to the full range of tools and processes they would use in a real case. Towler et al. (2018) state that both types of studies are necessary to truly understand the expertise underlying forensic practitioner decision making. The experiments in Chapter 6 tested the perceptual skill of three trained forensic face examiners on a series of quick decision face-matching tasks. The examiners consistently outperformed controls at the group level, however, not all examiners were statistically superior to controls at an individual level. The examiners did make fewer high confidence errors, but in all other regards performed similarly to high performing controls. Although face examiners are believed to rely on a feature-based face-matching strategy in operational casework (Towler et al., 2021), it is likely that when matching faces quickly they use a combination of featural and configural face-matching processes (Growns & Martire, 2020b), meaning that their perceptual skill may not reflect their operational accuracy.

The procedures used by forensic face examiners in operational casework are more complex than quickly matching two side-by-side faces. Current best practice recommends that forensic face examinations are conducted within an ACE-V framework (analyse, compare, evaluate – verify) (European Network of Forensic Science Institutes, 2018). First, examiners analyse the face images to determine the quality, quantity and type of facial feature detail visible. Next, the results of the analyses are compared between the face images. Then the examiner evaluates the strength of their observations as a level of support for whether the face images are the same person or different people, with the findings being independently verified by a second examiner using the ACE process (for a more detailed overview and case example see Moreton, (2021). Therefore, whilst perceptual skill tests are informative for understanding the face-matching abilities of individual examiners, they are far removed from how face examiners work and may not provide a realistic indication of examiner accuracy in forensic casework. Phillips et al. (2018) presented results from the largest study to date of face examiner operational accuracy and found that when matching faces within casework conditions, examiners were statistically superior to student controls and a matched control group of fingerprint examiners. However, there were large individual differences in examiner performance, which is highly concerning. A potential limitation of this study is that the examiners had to submit their responses individually, whereas current practitioner guidance recommends that all forensic face examinations are independently verified by a second examiner (European Network of Forensic Science Institutes, 2018), meaning that the results of operational face-matching cases are often team decisions from more than one examiner. Tests of the operational accuracy of fingerprint examiners found that independent verification substantially reduced the number of errors compared to decisions made by a single examiner (Ulery et al., 2011). Therefore, the absence of verification means that the performance of face examiners presented by Phillips et al. (2018) may be an understatement of true operational accuracy.

The aim of the current study was to compare the operational accuracy of individual face examiners and face examiner teams on a face-matching task conducted in casework conditions. Examiner performance was benchmarked against a sample of untrained police controls. As well as comparing performance at the group level, examiner performance was also evaluated using individual case analysis, to better understand variation between different individual examiners and examiner teams, as recommended in super recogniser research (Noyes et al., 2017).

Forensic face examiners typically present their conclusions as a level of support for whether the images under examination depict the same person or different people (European Network of Forensic Science Institutes, 2018). This study compared the use of support levels by forensic examiners and controls when making face-matching decisions, to understand if examiners were more adept at using support levels and whether individual examiners differed from examiner teams in their use of the levels. Finally, given the effectiveness of human-algorithm fusion on face-matching performance observed in Chapter 7, decisions from individual examiners and examiner teams were fused with the similarity scores of an automated facial recognition algorithm to evaluate the benefits of fusion with expert groups.

## *8.2. Method*

### 8.2.1. Participants

The face-matching task was distributed by the European Network of Forensic Science Institutes (ENFSI) to forensic face-matching agencies as part of the 2018 ENFSI facial image comparison proficiency test for forensic face examiners. Results were received from 21 individual face examiners and 18 examiner teams working for 27 difference forensic face-matching agencies, including police departments, national forensic laboratories, commercial forensic providers and universities. Due to anonymisation of the data by ENFSI, demographic information about the examiner participants are not available. A control group of 65 police officers and staff from a UK police force (20 female, median age 40, age range 25 – 67) participated in the study online via the Qualtrics platform. The same proprietary DCNN facial recognition algorithm used in Chapter 7 was also tested as part of the study.

### 8.2.2. Materials

The face-matching task used in the study consisted of 20 face pairs. Of the 20 face pairs 13 were matching pairs and seven were non-matching pairs. Six of the face pairs depicted females. The individuals shown in the images ranged from 32 to 63 years of age and were of Spanish and Mediterranean descent. The images ranged in resolution from 178 pixels wide and 219 pixels high to 556 pixels wide and 716 pixels high. The images depicted front-facing individuals with variation in pose, expression and illumination. The test materials were designed and despatched to examiner participants by the Comisaría General de Policía Científica of the Spanish National Police. Examiner results and test materials were provided anonymised by the ENFSI Digital Imaging Working Group.

### 8.2.3. Procedure

All human participants responded to the face-matching test using an 11-point Likert scale to denote a level of support for whether the images were a match or a non-match. The support scale ranged from *"-5 extremely strong for the proposition that the images are not the same person"* to *"+5 extremely strong support for the proposition that the images are the same person"*, with the centre point of the scale denoting *"0 support for neither proposition".* In order to incorporate no support decisions in the analysis performance was calculated using the area under the receiver operating characteristic curve (AUC). AUC does not stipulate a criteria for match and non-match decisions, thus allowing for the inclusion of no support decisions. AUC also allows for comparison of the algorithm scores and human-algorithm fusion scores within the analysis.

Forensic face-matching agencies participated in the proficiency test using their standard operating procedures and software. The proficiency test was designed to closely replicate typical face examiner casework and was expected to take two to three days per individual examiner. Agencies were permitted to submit responses from individual examiners or an examiner team, specifying in their submission the type of response. Current practitioner guidance recommends that forensic face examiners use a feature-based morphological approach to match faces and work within an ACE-V framework (European Network of Forensic Science Institutes, 2018; Facial Identification Scientific Working Group, 2019a), however, due to the ENFSI proficiency test being a black box test the procedures used by the examiner participants are not known.

Control participants completed the face-matching task online using Qualtrics in the participants' place of work. Prior to completing the trials participants consented to take part in the study. Participants were then shown face images in side-by-side pairs and asked to

respond whether the faces were a match or a non-match using the 11 point Likert scale of support. The task was self-paced. The face pairs were also processed by the facial recognition algorithm, resulting in a similarity score. Comparing two identical images would result in a score of zero, as the faces become more dissimilar the similarity score increases, therefore lower scores are more indicative of a match.

The study received a favourable opinion by the ethical committee of the Open University.

## *8.3. Results*

### 8.3.1. Performance

Summary statistics of AUC scores by group are shown in Table 49. AUC scores for controls were normally distributed (W = .98, p = .801), however neither individual examiners (W = .90, p = .037) or examiner teams (W = .81, p = .002) were normally distributed. Both examiner groups outperformed controls at the group level, with examiner teams having the highest average AUC score (Figure 53). A Kruskal-Wallis test revealed a significant difference between groups with a large effect size ($\chi^2(2, 86) = 59.76$, $p < .001$, $\varepsilon^2 = .58$). Post hoc pairwise tests using Dunn's test with Bonferroni correction revealed a statistically significant difference between controls and individual examiners ($p < .001$) and controls and examiner teams ($p < .001$). There was no significant difference between individual examiners and examiner teams ($p = .255$).

*Table 49 – Summary statistics of AUC scores by group for ENFSI test images*

| ENFSI Test AUC | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
|---|---|---|---|---|---|---|
| Controls (N = 65) | .752 (.106) | .753 | .456 | .692 | .819 | .967 |
| Examiner individual (N = 21) | .914 (.065) | .907 | .720 | .885 | .965 | 1 |
| Examiner teams (N = 18) | .973 (.030) | .989 | .923 | .946 | .999 | 1 |

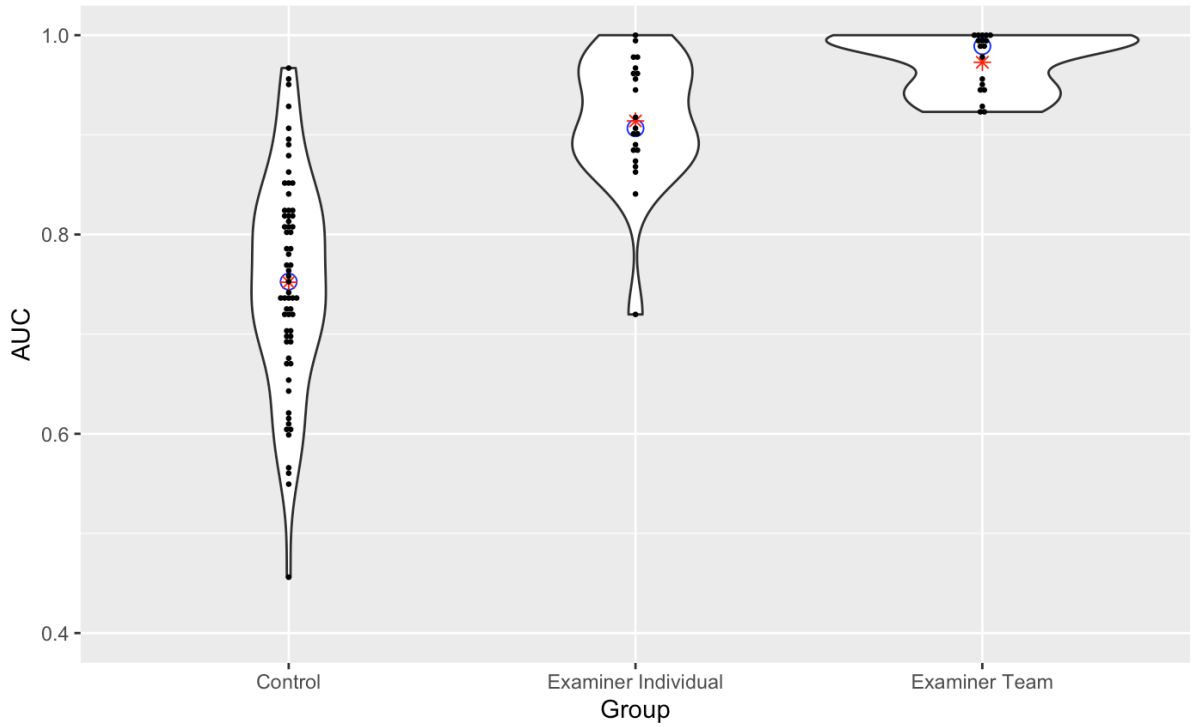


*Figure 53 - AUC scores for ENFSI test by group (blue circle represents median group score and red star is the mean group score*

To better understand the individual differences in examiner performance, each individual examiner and examiner team were compared to the control sample using individual case analysis (Crawford et al., 2010). Individual analysis revealed that 15 out of 18 examiner teams (83%) out performed controls at a statistically significant level using a one-tailed test and 11 out of 18 examiner teams (61%) were significantly superior to controls using a two-tailed test. The performance of the three remaining teams surpassed the average control score. Only 9 of the 21 individual examiners (43%) outperformed the controls at a statistically significant level using a one-tailed test and 5 out of 21 individual examiners (24%) outperformed controls using a two-tailed test.

### 8.3.2. Types of errors

Chapter 6 demonstrated that, on a test of perceptual face-matching skill, when individual examiners did make errors they did so with lower confidence than control participants. The purpose of the next analysis was to understand if examiners were equally cautious using support levels when carrying out face examinations, particularly when making an error. The proportion of errors at each support level (weak support, support, strong support, very strong support and extremely strong support) was calculated for controls, individual examiners and examiner teams. The distribution of error proportions by support level are shown in Figure 54.

As a group, examiner teams made the smallest number of errors, with no errors made at ‘very strong’ and ‘extremely strong’ levels of support. Only two teams made errors with ‘strong’ levels of support. Individual examiners were less cautious, with some individuals making errors with ‘strong’, ‘very strong’ and ‘extremely strong’ levels of support.

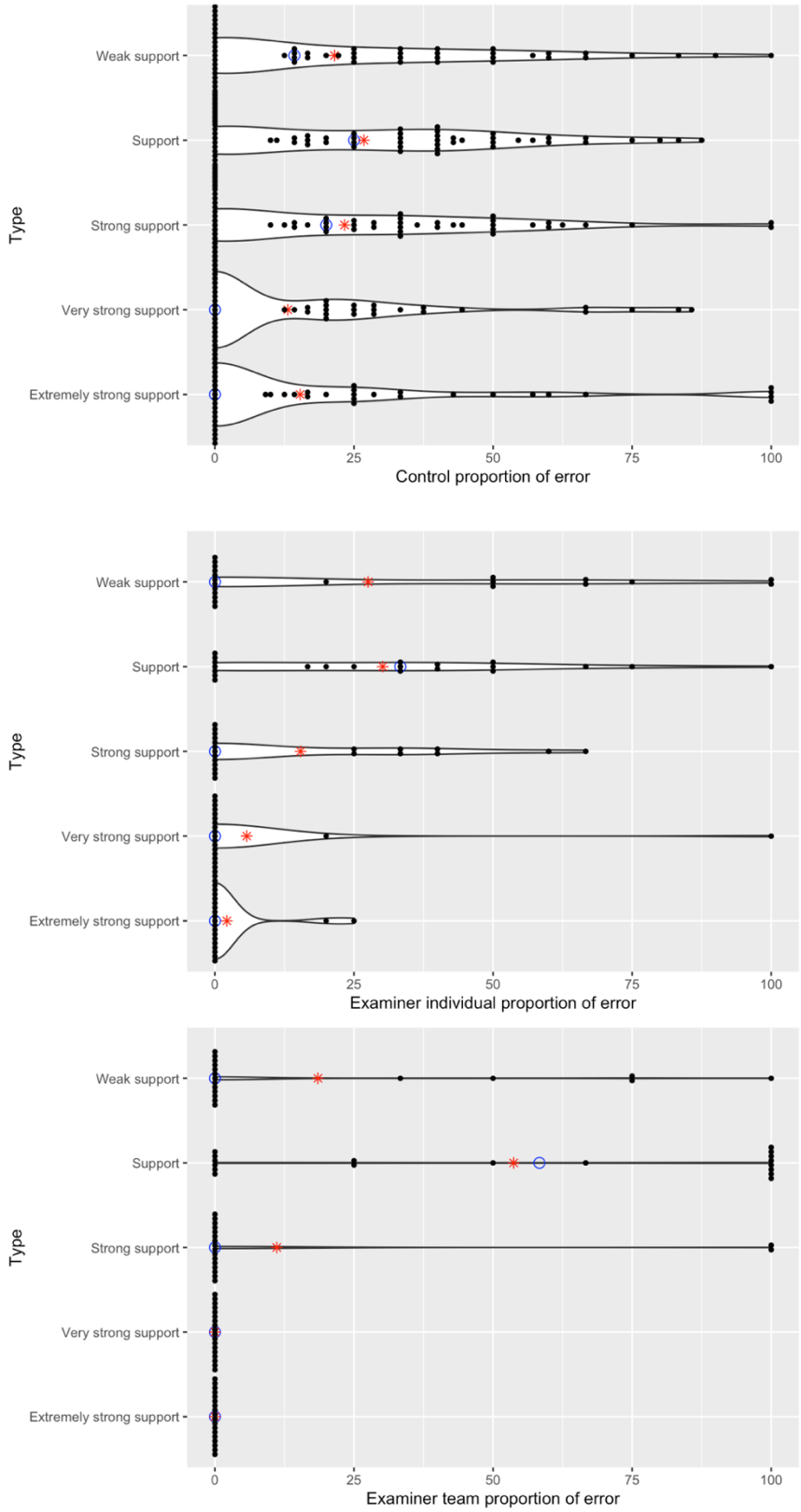


*Figure 54 – Distributions of proportions of error rate by support level for controls, individual examiners and examiner teams*

The proportions of error for each support level were compared between the three groups using the Kruskal-Wallis test, with post hoc pairwise tests using Dunn's test with Bonferroni correction for significant group effects. There was no significant difference between groups for 'weak support' errors ($\chi^2(2, 86) = 1.32$, $p = .516$, $\varepsilon^2 = .01$) or 'support' errors ($\chi^2(2, 86) = 4.19$, $p = .012$, $\varepsilon^2 = .04$). A significant effect was found for 'strong support' errors with a moderately small effect size ($\chi^2(2, 86) = 8.96$, $p = .011$, $\varepsilon^2 = .09$). For 'strong support' errors a statistically significant difference was observed between examiner teams and controls ($p = .011$) but not between examiner teams and individual examiners ($p = .549$) or controls and individual examiners ($p = .506$). Differences between groups for 'very strong' support errors were statistically significant with a medium effect size ($\chi^2(2, 86) = 12.78$, $p = .002$, $\varepsilon^2 = .12$), post hoc test revealed a difference between examiner teams and controls ($p = .005$) but not between examiner teams and individual examiners ($p = 1$). The difference between controls and individual examiners reached significance ($p = .049$). Finally, a significant difference with a medium effect size was found for 'extremely strong support' errors ($\chi^2(2, 86) = 13.56$, $p = .001$, $\varepsilon^2 = .13$), with a statistically significant difference between examiner teams and controls ($p = .005$), individual examiners and controls ($p = .031$) but not between examiner teams and individual examiners ($p = .849$). These results demonstrate that examiner teams are the least likely to make errors with high levels of support, followed by individual examiners and then controls.

Participants could also make 'no support' decisions for face pairs, when they felt unable to decide whether the face images were a match or non-match. Figure 55 shows the proportions of 'no support' decisions as a percentage of all decisions for controls, individual examiners and examiner teams. There was no significant difference in the proportions of 'no support' decisions between groups ($\chi^2(2, 86) = 1.18$, $p = .553$, $\varepsilon^2 = .01$), demonstrating

that the lower error rates of the examiner groups was not due to participants responding with a disproportionately high number of no support decisions.

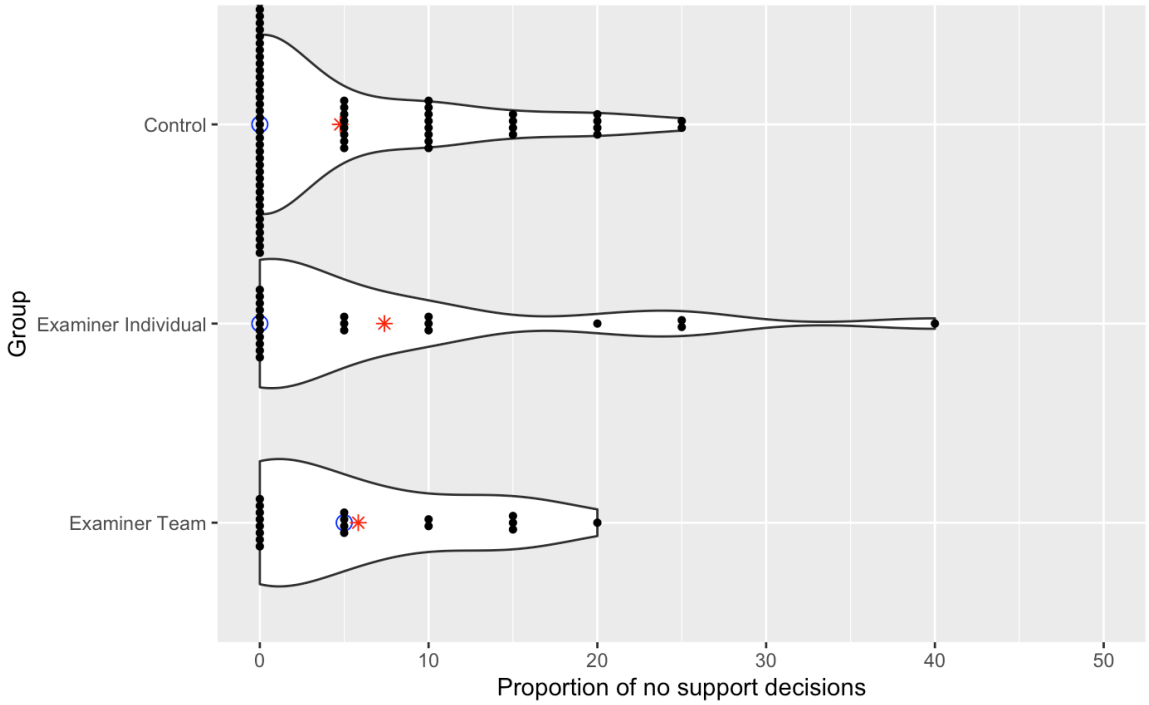


*Figure 55 – Distributions of proportion of no support decisions by group for ENFSI test*

### 8.3.3. Consistency of support levels

The aim of this stage of the analysis was to understand if examiners were more consistent than controls in their use of the support levels. Spearman's rho ($r_s$) was used to calculate correlation coefficients for support levels between participants within each group. Summary statistics of all Spearman's rho values (Table 50) show that controls were the least consistent group, with some participants being negatively correlated. Individual examiners were more consistent than controls in their use of support levels and examiner teams were the most consistent, at the group level.

*Table 50 – Summary statistics for correlation coefficients of decision rating agreement by group*

| Correlation coefficients of decision ratings | | | | | | |
|---|---|---|---|---|---|---|
| Spearman's rho | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
| Controls (N = 65) | .297 (.225) | .315 | -.615 | .160 | .454 | .868 |
| Examiner individual (N = 21) | .680 (.136) | .694 | .249 | .577 | .787 | .947 |
| Examiner teams (N = 18) | .793 (.078) | .802 | .507 | .753 | .843 | .942 |

The distributions of median $r_s$ values for each individual participant per group are shown in Figure 56. A Kruskal-Wallis test of median $r_s$ values revealed a significant difference between the groups with a large effect size ($\chi^2(2, 86) = 75.05$, $p < .001$, $\varepsilon^2 = .73$). Post hoc pairwise tests with Bonferroni adjusted p-values found that the differences between controls and individual examiners ($p < .001$) and controls and examiner teams ($p < .001$) were both statistically significant. Although examiner teams were more consistent than individual examiners at the group level, this difference was not statistically significant ($p = .317$)

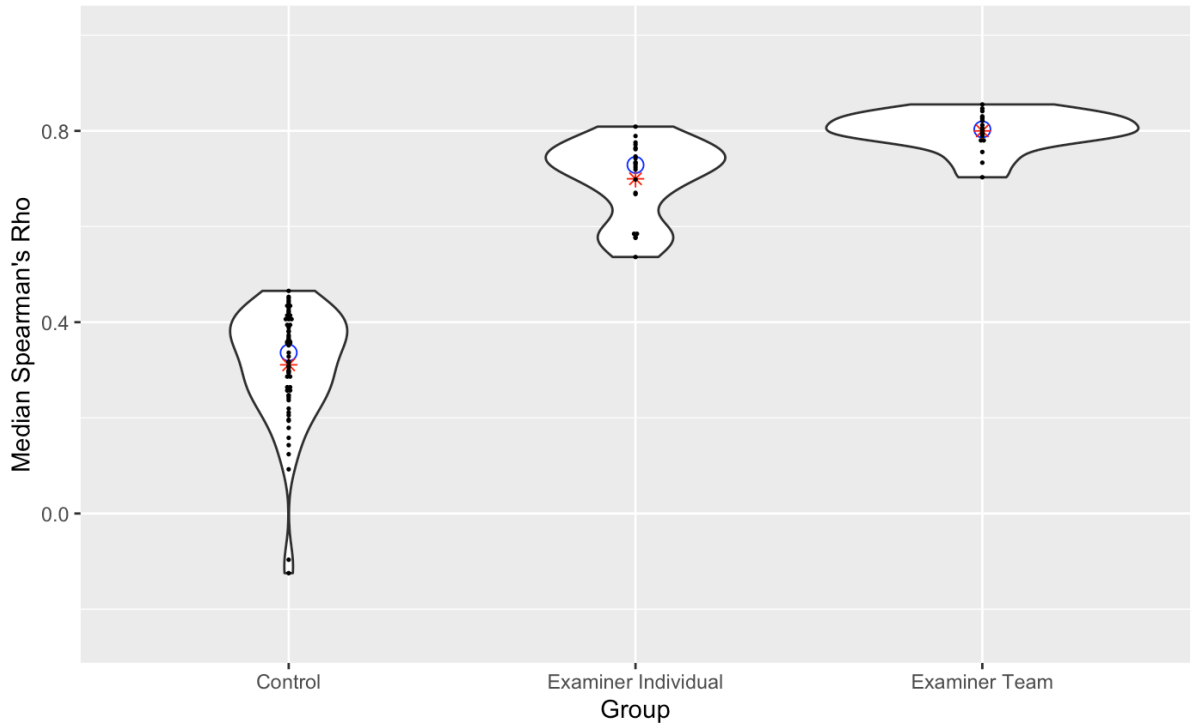


*Figure 56 – Median correlation coefficients of decision rating agreement for all human participants by group*

### 8.3.4. Fusion

Face pairs from the ENFSI test were processed by the DCNN algorithm used in Chapter 7 to allow a fusion analysis using control, individual examiner and examiner team decision ratings. The fusion analysis followed the procedure documented in Section 7.1.3. Summary statistics of AUC scores for the algorithm and fused groups are shown in Table 51. The algorithm outperformed the majority of controls, however individual case analysis revealed this was not significant using a one-tailed test ($t(64) = 1.61$, $p = .057$). Compared to examiners, the algorithm performed similarly to the average individual examiner, however all examiner teams performed at the same level or above the algorithm. The distributions of AUC scores pre- and post-fusion are shown in Figure 57. Fusion of human decision ratings and algorithm similarity scores resulted in increases in performance for all groups. The distribution of AUC scores for fused controls was similar to that of individual examiners

without fusion. Fused individual examiners had a similar range of AUC scores to unfused examiner teams but with more participants at ceiling, however not all individual examiners surpassed the algorithm AUC score after fusion. For examiner teams the AUC scores of all fused teams were approaching ceiling, limiting the extent of observable improvement for this group.

*Table 51 – Summary statistics of AUC scores for the algorithm, unfused and fused groups on the ENFSI test images*

| ENFSI Test AUC | Mean (SD) | Median | Min | 1st Quartile | 3rd Quartile | Max |
|---|---|---|---|---|---|---|
| Algorithm | .923 | - | - | - | - | - |
| Controls (N = 65) | .752 (.106) | .753 | .456 | .692 | .819 | .967 |
| Fused controls (N = 65) | .915 (.058) | .923 | .703 | .890 | .956 | 1 |
| Examiner individuals (N = 21) | .914 (.065) | .907 | .720 | .885 | .965 | 1 |
| Fused examiner individuals (N = 21) | .967 (.035) | .967 | .879 | .956 | 1 | 1 |
| Examiner teams (N = 18) | .973 (.030) | .989 | .923 | .946 | .999 | 1 |
| Fused examiner teams (N = 18) | .991(.011) | 1 | .967 | .978 | 1 | 1 |

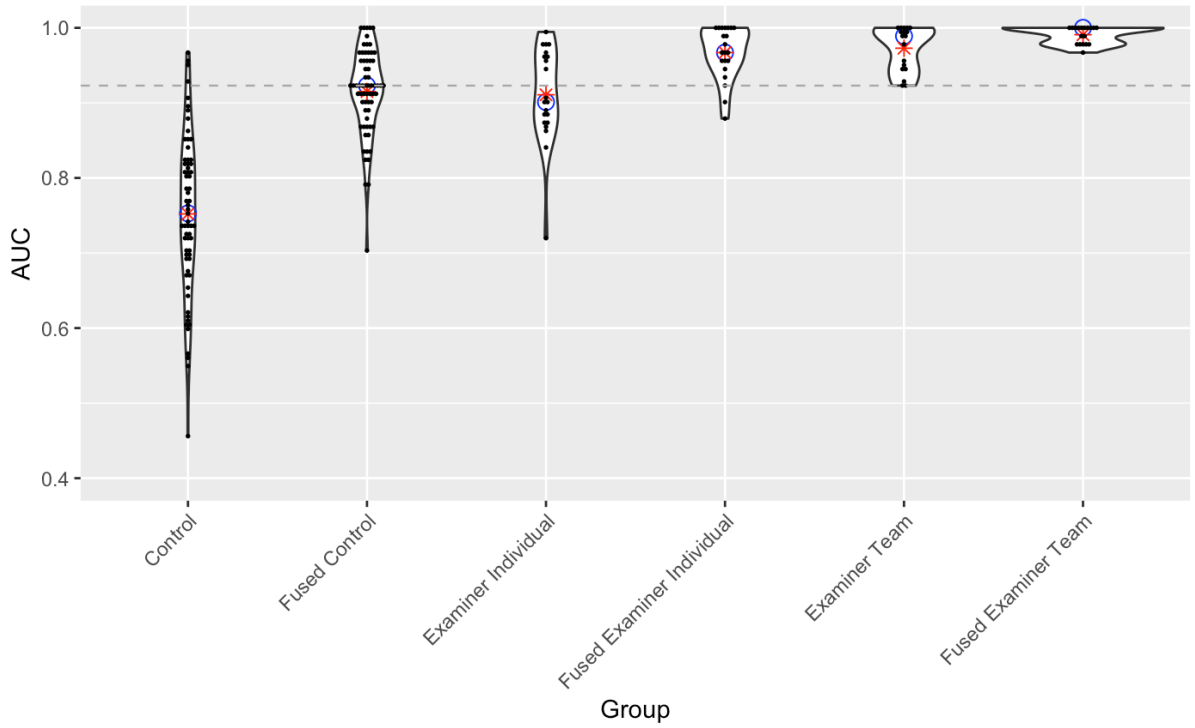


*Figure 57 – AUC scores by group for the ENFSI test images (grey dashed line represents algorithm score)*

Paired Wilcoxon signed-rank tests with Bonferroni adjusted p-values revealed statistically significant increases in AUC scores after fusion for controls ($V = 0$, $p < .001$, $r = .869$), individual examiners ($V = 4.5$, $p < .001$, $r = .832$) and examiner teams ($V = 4.5$, $p = .013$, $r = .638$). A Kruskal-Wallis test of AUC scores for the fused groups revealed a significant difference with a large effect size ($\chi^2(2, 86) = 39.39$, $p < .001$, $\varepsilon^2 = .38$). Post hoc Dunn's tests with Bonferroni adjusted p-values confirmed that the difference between fused individual examiners and fused controls ($p < .001$), fused examiner teams and fused controls ($p < .001$) were statistically significant, but not between fused examiner teams and fused individual examiners ($p = .218$). Individual case analyses of fused examiner teams revealed that all fused teams were statistically superior to controls using a two-tailed test. For fused individual examiners 18 out of 21 (89%) were statistically superior using a one-tailed test and 13 out of 21 (62%) were statistically superior using a two-tailed test.

## *8.4. Discussion*

Individual examiners and examiner teams were superior to controls at the group level on an operationally realistic forensic face-matching task. However, similar to previous studies, there were individual differences in examiner performance. Although there was no statistically significant difference between individual examiners and examiner teams at the group level, individual case analysis revealed that 83% of examiner teams were statistically superior to controls versus only 43% of individual examiners. Thus, examiner teams outperformed individual examiners and were the highest performing group in the study overall. Both examiner groups were more cautious than controls and were much less likely to make errors with strong support levels and examiner teams were the least cautious and made no errors with very strong or extremely strong levels of support. Examiner teams were also the most consistent in their decision making. These results support the current recommended best practice for independent verification by multiple examiners in forensic face examination casework. However, it should not be overlooked that some examiner teams were not statistically superior to controls and consistency in decision making was far from perfect for many examiner teams. It is important for future research to identify and address the underlying causes of these differences in examiner operational accuracy, whether it be due to training (as suggested in Chapter 4), procedures, underlying differences in perceptual skill or another as yet unknown factor. To do so requires future studies of operational accuracy to be white box tests that will allow researchers to unpick what makes a top-performing examiner team exceptional. White box testing has been used to understand variation in fingerprint examiner decision making (e.g. Ulery et al., 2016), but has as yet been unexplored in face examiner research.

The final part of this study examined the effects of fusion using human decision ratings and algorithm similarity scores. As in Chapter 7, fusion introduced significant gains in face-matching performance for all human groups. Prior to fusion all examiner teams performed at or above the AUC score of the algorithm, with fusion resulting in a significant improvement in performance for all fused teams compared to fused controls. Fusion also improved the performance of individual examiners, with 89% of fused individual examiners being statistically superior, a twofold increase over unfused individual examiners. Having multiple examiners work on a single case as a team is very resource intensive. For agencies with limited numbers of examiners, incorporating an algorithm into the examination process could be a cost-effective way of improving operational accuracy.

For examiner teams, collaboration with an algorithm would still appear to be worthwhile, in order to achieve the highest levels of performance. The use of examiner-algorithm teams presents an exciting area for further research that is only just beginning to be explored. Macarulla Rodriguez et al. (2020) have recently used three high-performing open-source facial recognition algorithms to generate likelihood ratios for face-matching decisions. Likelihood ratios are a common approach for evaluating the weight of forensic evidence in the courtroom (Aitken, 2008). They found that incorporating algorithm likelihood ratios with examiner derived likelihood ratios reduced the number of non-match errors for low quality face pairs. As well as white box testing, research should look to progress the integration of forensic examiner and algorithm expertise in face-matching, to ensure the reliability and validity of forensic face-matching evidence used in police investigations and the courtroom.

# 9. General discussion

## 9.1. Summary of findings

The research presented in this thesis explored four potential sources of face-matching expertise, namely training, superior face-matching ability, forensic facial examination and automated facial recognition algorithms, investigating the strengths and limitations of each in a series of experiments. The thesis also considered how different sources of expertise could be combined and implemented in applied face-matching systems.

Chapter 4 presented the results from an international survey of applied face-matching training, giving a snapshot of how agencies deliver training to operational face-matching professionals and providing the groundwork for further empirical work. The survey found that training practices between different agencies were highly diverse in terms of delivery methods, duration and content. Results also demonstrated that the use of evidence-based training practices, such as providing feedback on face-matching decisions, was varied, highlighting a potential disconnect between the research and applied face-matching communities. Of the agencies surveyed, 40% provided training for facial reviewers of five days or less, which is particularly concerning given the limited effectiveness of short face-matching training courses shown in the literature (Towler et al., 2019; Woodhead et al., 1979). Chapter 5 built upon previous work by Towler et al. (2019), empirically evaluating a two-day professional face-matching training course, further demonstrating that trainee face-matching accuracy did not improve after training, but also finding unexpected changes in the response bias of low performers post training.

Arguably the inherent limitations of short training courses suggests the focus should instead be on the perceptual skill of face matchers, where longer term training mentoring is not possible or feasible. This approach was explored in Chapter 6. Untrained superior face matchers were selected from a small pool of candidates and their performance was compared to trained forensic face examiners on a quick decision face-matching task. The study then used the wisdom of crowds paradigm to combine decisions from superior face matchers and face examiners, demonstrating that independent group decision making (the wisdom of crowds) gave significant gains in accuracy for both untrained superior matchers and trained face examiners. This performance gain was most pronounced for the untrained superior matchers, possibly due to the greater diversity in face-matching strategies used by this group. Crowd sourcing face-matching decisions also reduced the prevalence of high confidence errors by superior face matchers, a further benefit for applied face-matching.

The concept of independent group decision making was further explored in Chapter 7 by fusing human and algorithmic face-matching decisions, following the procedure used by Phillips et al. (2018). Human algorithm fusion was highly effective at improving face-matching accuracy, replicating findings from previous studies (O'Toole et al., 2007; Phillips et al., 2018). Fusion was most effective when the algorithm performed well, with gains seen across the full range of human performance. There was also some evidence for diversity in decision making between humans and the algorithm, as human and algorithm performance varied on different face pairs. Fusing the decisions of top human performers and algorithm similarity scores led to performance gains that surpassed both the maximum performance of individual humans and that of the algorithm.

Finally, Chapter 8 assessed the operational accuracy of professional forensic face examiners in an international face-matching proficiency test. Individual face examiners and face examiner teams were significantly superior to controls at the group level. Individual

case analysis revealed that a higher proportion of examiner teams were statistically superior compared to individual examiners and that examiner teams were also more conservative when assigning levels of support to their findings when making an error. Fusing individual examiner decisions with algorithm similarity scores resulted in performance comparable to examiner teams, and fusing examiner team decisions with the algorithm lead to almost ceiling levels of performance, further demonstrating the potential advantages of combining top performing humans with accurate facial recognition technology.

The following sections discuss the findings for each of the four sources of face-matching expertise in detail, highlighting the contribution of the current research to both the academic literature and applied working practices.

### 9.1.1. Training

Face-matching practitioner working groups advocate training for operational face-matching practitioners and have produced international guidance and best practice documents that provide recommendations for the content of such training (European Network of Forensic Science Institutes, 2018; Facial Identification Scientific Working Group, 2012b). An international survey of face-matching training (Chapter 4) found that, despite the existence of guidance documentation, there was substantial variation in the content, delivery and duration of training being delivered by different face-matching agencies. This study is believed to be the first to demonstrate the extent of diversity in applied face-matching training practices. Training duration was particularly diverse, with face reviewer training ranging from less than one day to up to 12 months. For face examiners training ranged from 2 weeks to five or more years. That training procedures were being inconsistently delivered by different face-matching agencies around the world is concerning, given that training is considered a source of face-matching expertise for professional practitioner groups. It

would appear that current best practitioner guidance is either lacking in sufficient detail to produce consistent training approaches or, if it does, is being ignored. Martire & Kemp (2016) produced a helpful and informative article aimed to assist forensic and applied practitioners in the design of human performance testing, with advice on participant recruitment, stimuli selection and test design, from the perspectives of cognitive scientists and psychologist. A new, similar article addressing the creation of effective training strategies would be highly beneficial to the applied face-matching community, particularly as survey results demonstrated the varied use of evidence-based training practices by face-matching agencies.

Short training courses of three days or less provide limited, if any, improvement in face-matching accuracy (Towler et al., 2019; Woodhead et al., 1979). Concerningly, over 40% of agencies participating in the survey provided training that was 5 days or less in duration. The results from Chapter 5 support previous research in demonstrating that short face-matching training courses are not, on their own, a reliable source of face-matching expertise. Following a two-day professional face-matching training course, 21 police trainees showed no overall improvement in face-matching accuracy. The study presented in Chapter 5 used similar pre- and post-training test stimuli and evaluated a course that was similar in content and delivery to Towler et al. (2019). Interestingly, where Towler et al. did find a slight improvement in accuracy on certain stimuli after completion of the course (face pairs from the GFMT and the MFMT), this was not replicated in the current study. This highlights that not only are short training courses very limited in improving face-matching accuracy, but where improvements do occur, they are inconsistent and hard to replicate.

Consistently superior performance on a task is one of the primary hallmarks of task-specific expertise (Academy of Social Sciences in Australia Inc., 2020), but Chapter 5 demonstrated that a short training course did not result in superior face-matching performance. Another

hallmark of expertise, particularly in forensic pattern-matching disciplines, is knowing when not to make a decision (Towler et al., 2018). Such expertise requires an understanding of when a pattern or stimuli is too poor or contains insufficient information to make a decision. There is some evidence that forensic face examiners are more cautious when matching low quality CCTV-style images and are less likely to make errors with high confidence compared to novices (Norell et al., 2015). It is, therefore, important to understand whether short training courses can be used to develop this type of expertise. As well as measuring overall face-matching accuracy post training, the experiment in Chapter 5 also investigated if there were any changes in face-matching confidence after training. Similarly to overall accuracy, the results of the experiment found no consistent change in the use of confidence ratings when making face-matching decision after training. After training, trainee confidence ratings were also no better calibrated to face-pair difficulty than controls. This demonstrates that, in addition to not improving overall face-matching accuracy, the training course did not result in trainees becoming more cautious in their face-matching decisions, further highlighting the limitations of short face-matching training courses

Concerningly, there was an unexpected change in face-matching behaviour observed post-training. Trainees who were initially poorer at face matching had a significant shift in response bias after training, showing an increased tendency to respond match rather than non-match. This significant shift in response bias was not accompanied by a significant improvement in sensitivity. As a result, lower performers were more likely to correctly identify matches, (an increased hit rate), but were also at risk of increasingly misidentifying non-matches, (an increased false alarm rate). Criterion shifting has been researched in recognition memory, showing large individual differences in how people shift their criterion on old/new recognition tasks (Layher et al., 2020). In certain scenarios criterion shifting can be advantageous for improving performance, for example, in a recognition task where a

target occurs frequently, shifting to a liberal response bias may improve overall performance on the task. However, in order for a criterion shift to improve performance the base rate probabilities of target prevalence must be known and understood. Even when such information is available individuals often do not shift criterion effectively (Aminoff et al., 2012). In the current study the mechanism causing the criterion shift in low performers is not readily apparent. Participants were not provided with any information concerning match and non-match prevalence, hence the change in response bias was not derived from an understanding of base rate probabilities. This is also reflected in the fact that the sensitivity of low performers did not improve after training.

It is possible that the training course in question teaches strategies that encourage match responses by instructing trainees to break up the face and compare individual features. Studies of feature-based face-matching strategies are limited but results so far have shown that instruction in such strategies improved accuracy for matching face pairs, but not for non-matching pairs (Megreya & Bindemann, 2018; Towler, White, et al., 2017). However, if the training course does teach strategies that encourage match responses, it is not clear why only low performing trainees were affected. It may be the case that the lower performing trainees have less stable decision criteria than the high performers and are more susceptible to external influences that can shift criterion, as suggested by Gentry & Bindemann (2019). Caution is advised in generalising this interpretation to all face-matching training, given that the current study was unable to fully replicate the findings of Towler et al. (2019), this observation could also be similarly hard to replicate. However, the results from Chapter 5 raise concerns as to the unexpected consequences of training courses that have not been rigorously tested. In applied settings, a shift in decision criterion could have far-reaching consequences, from the wrongful identification of suspects to missing hostile imposters. Therefore, when designing training material an evidence-based approach should

be used with consideration given to the impact of training beyond simply improving performance.

Whilst there is now an established body of evidence that short training courses are largely ineffective at improving face-matching ability, there is a significant and unanswered question in the literature of whether longer term training develops face-matching expertise. Trained forensic face examiners have demonstrated expertise in face-matching tasks (Norell et al., 2015; Phillips et al., 2018; White, Phillips, et al., 2015), which is believed to be derived from training and professional experience rather than solely based on natural ability (Towler et al., 2021). The results from Chapter 4 found that examiner training was often 1 year or more in duration, was more likely to be delivered by one-to-one mentoring and included a detailed range of topics. It could thus be inferred that longer, more detailed training and mentoring is contributing to the enhanced expertise of examiner groups observed in the literature, as suggested by Towler et al. (2021). However, this assumption is yet to be empirically validated.

The need for further evidence-based investigation of current face-matching training practices is pressing. The face-matching community must also look to validate the efficacy of longer duration training courses that reflect current practice through longitudinal studies. Doing so will help to establish if long term training is a viable source of expertise in professional face-matching groups. This will, in turn, provide an evidence base for future best practice and guidance in applied face matching training.

### 9.1.2. Superior face matchers and crowd effects

Given the limited effectiveness of short training courses in improving face-matching ability, Chapter 6 explored the selection of individuals with superior face-matching ability. Pre-recruitment screening is an approach advocated by researchers as a means for identifying individuals who naturally possess exceptional face recognition and perception abilities, referred to as super recognisers (SRs) (Bobak, Dowsett, et al., 2016; Davis et al., 2016). Pre-screening is recommended for applied settings where high volumes of faces must be matched quickly, such as at the border (Bobak, Dowsett, et al., 2016). Selecting individuals with naturally superior face-matching and recognition abilities appears to be a logical approach for improving performance in applied face-matching tasks, however, there are a number of challenges to consider regarding the real-world implementation of such screening procedures. Firstly, abilities in different face perception subprocesses are both highly varied and diverse between different individuals, meaning that a high performer in one type of process (e.g. familiar face recognition) may not necessarily perform highly on another process (e.g. unfamiliar face matching) (Fysh et al., 2020). Secondly, applied face recognition and matching tasks also vary considerably depending upon the intended output and the role of the person carrying out the task (Moreton et al., 2019). Therefore, for screening procedures to be fit for purpose they should test for the types of skills required to carry out specific tasks and reflect the operational requirements of that task.

Another challenge in SR recruitment is finding such individuals. SRs with truly exceptional face recognition and matching abilities are believed to comprise only one to two percent of the general population (Academy of Social Sciences in Australia Inc., 2020), thus simply identifying SRs for operational deployment can be difficult in of itself. In an applied setting, if the number of potential candidates for a role is small it becomes decreasingly likely that an exceptional SR will be found. Dunn et al. (2020) proposed a potential solution to this

problem by developing an online screening tool that aimed to identify exceptional SRs in the general population. Given the low prevalence of exceptional face recognition and matching abilities, Dunn et al. advocated large scale online testing as a mean to find such individuals who could then undergo further task-specific testing. The findings from Dunn et al.'s study appear promising, but this approach requires access to technology, expertise and resources to carry out large scale online testing that may not necessarily be available to all operational agencies. There may also be other requirements for an operational role beyond face perception ability that could potentially rule out candidates found through online testing.

In light of these challenges, Chapter 6 explored three alternative approaches to improving face-matching performance on a quick decision face matching task. The first approach was selecting superior face matchers (SMs) using a small selection pool of operational police personnel (n = 28) who had not received any training in face matching, using a representative face-matching task with associated performance data from a larger sample of individuals. This approach aimed to replicate an applied setting where only a limited number of recruits for a face-matching role are available. The second approach was to test the performance of trained forensic face examiners (FEs) on the quick-decision face-matching task and compare their performance to the selected superior face matchers from the untrained group. The final approach explored combining the decisions of multiple superior matchers and face examiners using the wisdom of crowds.

In the SM selection procedure, there were no individuals from the selection pool who surpassed two standard deviations above the mean, which is the standard cut off for SR selection in the literature (Fysh et al., 2020), demonstrating the difficulty in identifying truly exceptional SRs from small samples. However, the top three individuals all performed at least one SD above the mean. At re-test, despite the selection face-matching task and re-

test face-matching task being strongly correlated ($r(31) = .74$, $p < .001$), all three selected SMs deteriorated in face-matching accuracy. The SMs were also varied in their response bias between the two tasks. This demonstrates that even when a screening tool is representative of the task it is selecting for, it may not be possible to identify consistently superior performers, particularly from a small pool of individuals.

Three FEs also completed the selection and re-test face-matching tasks. The FEs had significantly superior performance on both tasks at the group level, however this enhanced perceptual skill was largely due to an individual exceptional FE. The two remaining FEs varied in both performance and response bias across tasks, displaying similar behaviour to individuals in the SM group. Therefore, using trained face examiners for high volume, quick decision face-matching tasks also has inherent limitations. The primary difference between FEs and SMs appeared to be in their use of confidence decisions, with FEs making not making any errors with high levels of confidence, whereas SMs did.

Although selecting SMs from a small pool of candidates gave limited gains in accuracy at re-test, combining this selection procedure with a wisdom of the crowds approach was highly effective. By independently combining the decisions of multiple SMs performance increased significantly. Interestingly, wisdom of the crowds for SMs was more effective than for FEs, suggesting that the SMs were more diverse in their decision making, which contributed to gains in accuracy, whereas FE crowd performance was driven by the top performing FE in the crowd. SM crowds also did not make high confidence errors, highlighting a further benefit of this approach.

The results from Chapter 6 replicate those of Balsdon et al. (2018), signifying the need to use a combination of techniques, in this case pre-screening and wisdom of the crowds, to provide significant gains in face-matching performance in scenarios where it is not possible

to identify truly exceptional superior face matchers. The current study builds on these findings by demonstrating this approach with operational police personnel and including a group of trained forensic examiners for comparison. The study also found an additional benefit of reducing high confidence errors using the wisdom of crowds. Thus, combining selection procedures with group decision making would appear to be an effective procedure for ensuring accuracy and reducing high-confidence errors in applied face-matching settings.

### 9.1.3. Operational accuracy of forensic face examiners

Chapter 6 found limited gains from using trained forensic face examiners on high volume, quick decision face-matching tasks, but such a task does not reflect how face examiners match faces operationally. Typically, face matching by forensic face examiners is a lengthy, rigorous process involving multiple sequential steps taking hours or even days depending upon the complexity of the examination (Moreton, 2021). Current international best practice also recommends that multiple forensic face examiners should work on an examination in a process known as blind verification, in order to improve the reliability of the result (European Network of Forensic Science Institutes, 2018). Testing of the operational accuracy of forensic face examiners, using examination procedures, has so far been limited. Phillips et al. (2018) published the most comprehensive study of face examiner accuracy to date, however in the study examiners had to work individually rather than in small teams, as is current best practice.

The aim of Chapter 8 was to evaluate the performance of an international cohort of forensic face examiners on a challenging face-matching task and investigate whether performance varied between examiners who worked individually and those who worked in teams. Examiner performance was compared to a control group of police personnel who had not

received training in face matching. Both individual examiners and examiner teams were significantly more accurate than controls. Examiner teams were more accurate than individuals, but this was not significant at the group level, there was also overlap in performance between all three groups. Individual case analysis revealed that 61% of examiner teams were significantly superior to controls using a two-tailed test, whereas only 24% of individual examiners were superior, demonstrating an advantage for examiners who worked in teams. An analysis of the types of face-matching responses showed that, unlike individual examiners, examiner teams did not make any errors with 'very strong' or 'extremely strong' levels of support. Examiner teams were not only the most accurate group, they were also the most consistent in the levels of support given to their face-matching decisions.

The findings of Chapter 8 contribute to the research literature by comparing the performance of individual examiners and examiner teams. The findings support current best practice recommendations that forensic face examiners should carry out their procedures in teams rather than individually. However, it is important to note that examiner teams did make errors. Whilst high-level recommendations for examiner procedures exist, the specifics of how different examiners and teams operate is not well understood. As Chapter 4 revealed that training practices for examiners varied between different agencies, further research should investigate whether face examination procedures are similarly diverse using a combination of task analysis and white box testing.

### 9.1.4. Combining human and algorithm expertise

Chapters 7 and 8 explored the fusion of human face-matching decisions and algorithm similarity scores. The benefits of human-algorithm fusion in face matching were first demonstrated over 13 years ago (O'Toole et al., 2007). More recently the benefits of fusion have been demonstrated with state-of-the-art DCNN facial recognition algorithms and high performing super recognisers and forensic face examiners (Phillips et al., 2018). In practice, however, automated facial recognition systems and human operators typically work in a sequential fashion, with the algorithm being used to reduce a database of faces to a list of possible candidates, which are then reviewed by a human. Research has shown this approach can be error prone, even when conducted by trained operators (White, Dunn, et al., 2015). Thus, further research on how best to combine human and algorithm face-matching expertise will be beneficial for applied face-matching systems that include these two elements.

Chapter 7 investigated fusion of an algorithm's similarity scores with human decision ratings on face pairs that were challenging to humans and face pairs that were challenging to an algorithm. For human-challenging face pairs the algorithm performed highly, whereas for algorithm-challenging face pairs the majority of human participants outperformed the algorithm, indicating that the algorithm may be matching faces in a different way to the human participants. Exactly how deep learning algorithms match faces is not well understood and may vary between different algorithms (O'Toole et al., 2018), however, on tests comparing the face-matching capabilities of algorithms and humans, algorithms were observed to make mistakes that would be highly unlikely for a human observer. Hancock et al. (2020) intentionally morphed the ethnicity and gender of same face pairs and found that five commercial algorithms were still likely to say the faces were a match, whereas human observers rated the faces as being different individuals. Hancock et al.'s findings

suggest that the algorithms were able to ignore non-identity variation in faces. Other researchers have suggested that the large face datasets used to train algorithms means that the face representations made by algorithms can generalise across image-specific and non-identity variation (Hill et al., 2019). In the results from Chapter 7 it appears that this diversity in decision making, in addition to performance, is a contributor to the effectiveness of fusion, as fusing the top performing humans and the algorithm resulted in further gains in accuracy. In Chapter 8, fusing algorithm similarity scores and decisions from individual forensic face examiners and examiner teams was also effective, with fused individual examiners performing at the level of unfused examiner teams and fused examiner teams approaching ceiling levels of performance.

Independently fusing human face-matching decisions and algorithm similarity scores appeared to be the most effective way of improving face-matching performance explored in this thesis, making human-computer interaction in face-matching a promising area for further research. These results also have implications for applied settings. For quick decision face-matching tasks (e.g. passport control) utilising fusion techniques may allow less stringent selection criteria for human operators. This would allow agencies to recruit more individuals at a lower threshold of face-matching ability. For scenarios that require forensic face examination (e.g. expert evidence for presentation at court) if there are insufficient resources to deploy multiple examiners, teaming individual examiners with an algorithm could be an effective solution.

## *9.2. Practical recommendations*

In applied settings, the use of face-matching technology by governments and private companies is steadily increasing and with this rise, somewhat ironically, the need for human adjudication of the results from such systems also increases (Academy of Social Sciences in Australia Inc., 2020). Given the potential ramifications of an incorrect face-matching decision in high-risk applied settings, such as policing and security, it is important that the face-matching systems and procedures used operationally are both accurate and evidence based. The research presented in this thesis has been conducted using police and forensic practitioners with and without professional face-matching training and experience, in a series of experiments that are applicable to how face-matching is being conducted in real-world settings. Based on the findings of these experiments a number of recommendations are made for how face-matching should be conducted in two different applied settings, the first being quick decision face matching, (e.g. checking passports at the border) and the other forensic face examination.

### 9.2.1. Recommendations for quick decision face matching

In quick decision face-matching settings operators must make large numbers of face-matching decisions under time pressure and often alongside other tasks, such as checking the authenticity of an identity document (Stevens, 2021). Operators may also be working with automated systems and comparing multiple candidates to a facial image (Heyer et al., 2018). In such settings the detailed face-matching procedures used by forensic face examiners are not applicable or appropriate. It also may not be feasible from a resourcing perspective for operators to undergo lengthy training and mentoring to develop their face-matching expertise. Figure 58 provides an overview of recommendations for improving

accuracy in quick decision face-matching settings, in scenarios where an algorithm is available and where the operator works without an algorithm.

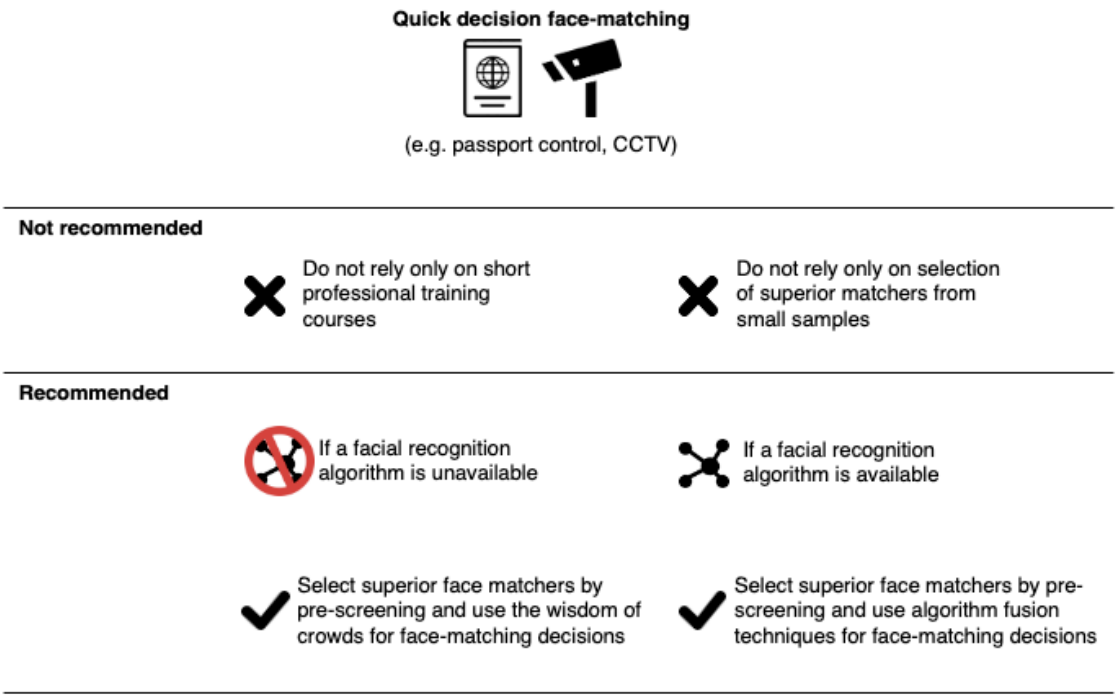


*Figure 58 – Recommendations for quick decision face-matching scenarios*

**Recommendation 1**: Agencies should not rely solely on short training courses to develop operator face-matching expertise. Instead, individuals with superior face-matching performance should be recruited using evidence-based selection tests that are representative of the real world face-matching tasks they will undertake operationally.

**Recommendation 2:** The decisions of multiple superior matchers should then be combined in a wisdom of the crowds approach to give further gains in accuracy.

**Recommendation 3:** In scenarios where a high-performing automated facial recognition algorithm is available the decisions of superior face matchers can be combined with the algorithm outputs in a similar fashion, reducing the number of humans required to make each decision and potentially freeing up resources.

### 9.2.2. Recommendations for forensic face matching

In forensic face examination, trained examiners apply rigorous face-matching procedures that are not subject to time constraints, however often the results must be interpretable and explainable to the end-user, such as an investigator or a court of law. Figure 59 provides an overview of recommendations for forensic face examination both when an algorithm is available and when it is not.

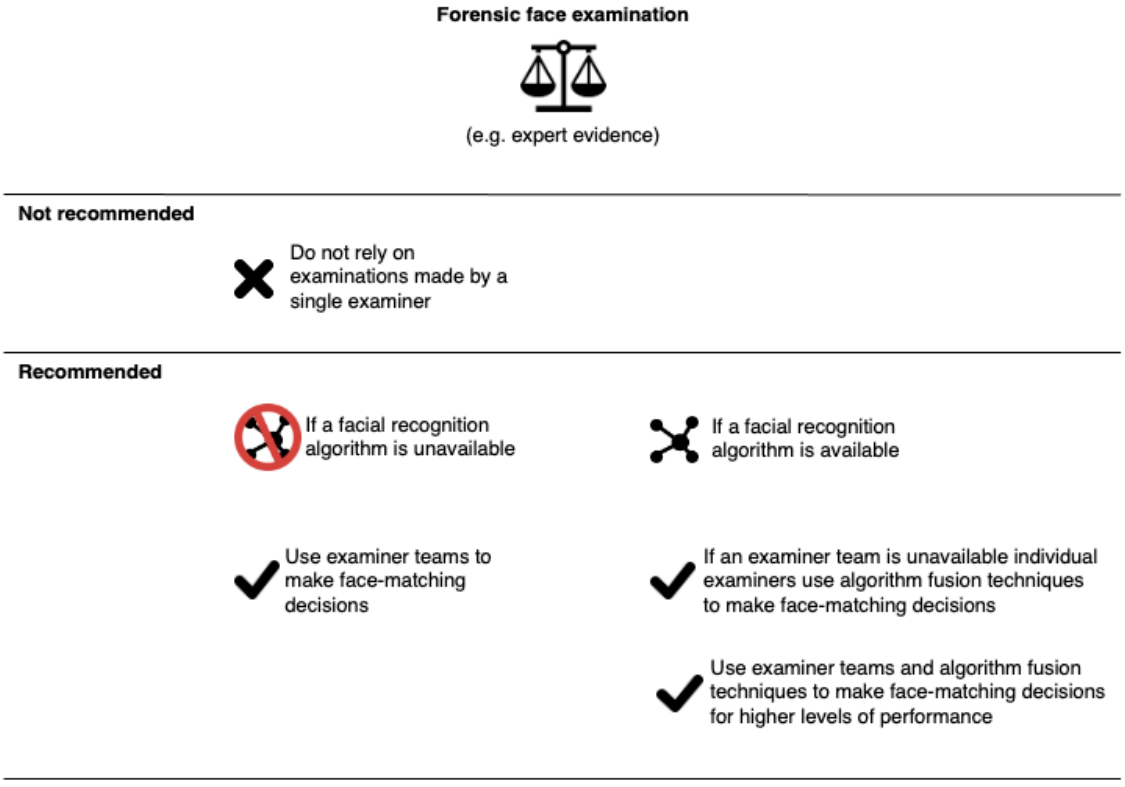


*Figure 59 – Recommendations for forensic face matching scenarios*

**Recommendation 4:** Given that the majority of individual face examiners performed within the range of controls it is recommended that forensic face examinations are not carried out by individual examiners, supporting current best practice for practitioners (European Network of Forensic Science Institutes, 2018). Instead, examinations should be conducted by teams of examiners.

**Recommendation 5:** If a high performing face-matching algorithm is available the decisions of individual examiners can be combined with algorithm outputs to improve accuracy. Where sufficient resources allow, combining a team of examiners with a high performing algorithm appears to provide the largest gains in face-matching accuracy.

Human-algorithm fusion would seem to be the most effective strategy for improving accuracy in both quick-decision face matching and forensic face examination, however it is important to highlight that further research is required to fully understand how this approach could be implemented in practice (see Section 9.3).

## *9.3. Limitations and future research*

The research presented in this thesis is derived largely from police and forensic personnel, some of whom are trained face examiners, using a wide range of face images including those that represent real-world case work conditions. However, there are a number of limitations to consider when interpreting the findings, which will hopefully be addressed in future research. Firstly, all of the experiments in this thesis consisted of comparing two faces. Operationally, the range of tasks that face-matching professionals undertake are much more diverse (Moreton et al., 2019), such as comparing one face to an array of candidate faces (Heyer et al., 2018) or searching for a target face in video footage (Mileva & Burton, 2019). Further research should test other face-matching paradigms., which will validate whether the conclusions from this thesis are applicable to different types of face-matching tasks, as improved performance on one task may not necessarily translate to improvements on another (Ramon et al., 2019a).

As well as being limited in the types of face-matching tasks tested, the experiments conducted in this thesis were completed in semi-controlled conditions and do not directly emulate the conditions that face-matching operators work in. In real world conditions there may be non-face information that can affect performance or bias decision making, such as identity document details (Feng & Burton, 2019) and age information (Robertson & Burton, 2021), as well as environmental factors and social interactions (Tummon et al., 2019). Ideally, further research into the performance of face-matching professionals should include experiments that aim to simulate the conditions under which the work is completed. Virtual reality technology may provide a promising and cost-effective avenue for conducting this type of research and has been explored by Tummon et al. (2019) to simulate face matching when checking passports at an airport.

A particular focus in this thesis has been given to the benefits of fusing human and algorithm face-matching decisions. Whilst this approach is highly effective in controlled experiments where the ground truth is known, there are a number of limitations in the current approach that require further research. For human-algorithm fusion to be effective, not only is it important to ensure that both the human and algorithmic elements are reliable and accurate. The design of such human-machine face-matching systems introduces additional challenges and considerations, particularly regarding interaction between the human and the algorithm. For example, even highly accurate algorithms may make unpredictable errors that would not be made by a human (Yang et al., 2020), such as declaring two faces match where one of the faces in a pair has been transformed to appear of a different sex or ethnicity (Hancock et al., 2020). Facial recognition algorithms are also susceptible to adversarial attacks that impair their functionality, through the introduction of perturbations in pixels that are imperceptible to human observers (Theagarajan & Bhanu, 2020). A human-algorithm face-matching system must be designed with sufficient safeguards to address such unexpected errors.

How algorithm results are presented within a graphical user interface have been observed to affect trust in autonomous image classifiers (Ingram et al., *in press*) and bias subsequent face-matching decisions made by human observers (Fysh & Bindemann, 2018; Howard et al., 2020). Therefore, consideration should also be given to how systems can garner appropriate levels of trust in the algorithm by the operator, ensuring that the operators does not distrust the algorithm (i.e. think that the algorithm cannot be relied upon) or over trust the algorithm and give undue confidence to its results.

In forensic disciplines in particular, practitioners have been reluctant to apply algorithms in casework, citing a wide range of concerns including anecdotal instances of algorithm failures and worries around scrutiny, oversight and quality control of algorithms (Swofford

& Champod, 2021). Another major issue in the implementation of algorithms in forensic casework is ensuring that any results are explainable and interpretable. This is particularly challenging as many algorithms are effectively black boxes with little understanding of how they actually work (Bollé et al., 2020). There is a growing body of research that aims to generate score-based likelihood ratios from image-matching algorithm outputs for a range of forensic disciplines, including forensic face-matching (see Jacquet & Champod, 2020), resulting in a weight of evidence (the likelihood ratio) that can, to some extent, be interpreted within the context of a forensic examination. Macarulla Rodriguez et al. (2020) recently found that using score-based likelihood ratios from a facial recognition algorithm aided forensic face examiners when matching low quality CCTV-style images. Work in this area is ongoing, but score-based likelihood ratios may be an effective approach for fusing forensic examiner and algorithm face-matching decisions in a manner that is explainable and interpretable.

Other areas for future research identified in this thesis also relate to forensic face examination, namely the need for longitudinal studies of examiner training in a similar manner to that used Searston & Tangen (2017b) for forensic fingerprint examiners. Longitudinal studies of training effectiveness will assist in understanding whether longer-term training and mentoring is a contributor to face examiner expertise, as proposed by Towler et al. (2021). There is also a need to move towards more open and transparent white box testing of face examiners, as has been done for fingerprint examiners (e.g. Ulery et al., 2016, 2017), to better understand why examiner performance varies on face-matching tasks.

## *9.4. Conclusion*

Face-matching is likely to continue as a means of identification in applied settings, despite longstanding concerns from academic researchers of human fallibility at the task. Much of the research to date has focussed on a single potential source of face-matching expertise, whether it be training, forensic examiners, super recognisers or algorithms. However, in real world face-matching systems multiple sources of expertise may be involved depending upon the task. This thesis has attempted to look across these various types of expertise and identify ways in which they can be combined to improve the overall accuracy of a face-matching system, as well as investigating the limitations of a given source of expertise (such as short training courses).

The recommendations from this thesis demonstrate that by combining different approaches, such as personnel selection with the wisdom of crowds, or face examiner teams with an algorithm, further gains in face-matching accuracy can be achieved. However, the appropriateness of the approach is dependent upon the conditions in which they are used, for example, employing forensic face examiners in high volume face-matching settings is not an effective use of their expertise. The current research provides a sound foundation for how face-matching performance can be effectively improved in applied settings. The research also identifies promising avenues for further research to develop these ideas, with the aim of developing applied face-matching systems that are evidence-based and consistently demonstrate expertise and superior performance.

# *10. References*


Academy of Social Sciences in Australia Inc. (2020). *Evaluating face identification expertise: turning theory into best practice*.

Aitken, C. (2008). Interpreting the Value of Evidence. *Royal Statistical Society, 9 Sept 2008*. http://www.rss.org.uk/PDF/Colin Aitken - 9 Sept 08.pdf

Alenezi, H. M., & Bindemann, M. (2013). The Effect of Feedback on Face-Matching Accuracy. *Applied Cognitive Psychology*, *27*, 735–753.

Alenezi, H. M., Bindemann, M., Fysh, M. C., & Johnston, R. A. (2015). Face matching in a long task: enforced rest and desk-switching cannot maintain identification accuracy. *PeerJ*, *3*, e1184. https://doi.org/10.7717/peerj.1184

Aminoff, E. M., Clewett, D., Freeman, S., Frithsen, A., Tipper, C., Johnson, A., Grafton, S. T., & Miller, M. B. (2012). Individual differences in shifting decision criterion: A recognition memory study. *Memory and Cognition*, *40*(7), 1016–1030. https://doi.org/10.3758/s13421-012-0204-6

Anderson, J. (1982). Acquisition of Cognitive Skill. *Psychological Review*, *89*(4), 369–406.

Andrews, S., Jenkins, R., Cursiter, H., & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *The Quarterly Journal of Experimental Psychology*, *68*(10), 2041–2050. https://doi.org/10.1080/17470218.2014.1003949

Ashcroft, J., Daniels, D. J., & Hart, S. V. (2004). *Education and Training in Forensic Science : A Guide for Forensic Science Laboratories, Educational Institutions, and Students*.

Austen, G. E., Bindemann, M., Griffiths, R. A., & Roberts, D. L. (2016). Species identification by experts and non-experts: Comparing images from field guides. *Scientific Reports*, *6*(August), 1–7. https://doi.org/10.1038/srep33634

Baer, J. (2015). The Importance of Domain-Specific Expertise in Creativity. *Roeper Review*, *37*(3), 165–178. https://doi.org/10.1080/02783193.2015.1047480

Ballantyne, M., Boyer, R. S., & Hines, L. (1996). Woody Bledsoe—His Life and Legacy. *AI Magazine*, *17*(1), 7–20. https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1207/1108

Balsdon, T., Summersby, S., Kemp, R. I., & White, D. (2018). Improving face identification with specialist teams. *Cognitive Research: Principles and Implications*, *3*(1). https://doi.org/10.1186/s41235-018-0114-7

Bartle, A., & Dellwo, V. (2015). Auditory speaker discrimination by forensic phoneticians and naive listeners in voiced and whispered speech. *The International Journal of Speech, Language and the Law*, *22*(2), 229–248. https://doi.org/https://doi.org/10.1558/ijsll.v22i2.23101

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A. K., Wills, H., & Richards, S. (2018). Applied screening tests for the detection of superior face recognition. *Cognitive Research: Principles and Implications*, *3*(1), 22. https://doi.org/10.1186/s41235-018-0116-5

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Portch, E., Murray, E., & Dudfield, G. (2019). The consistency of superior face recognition skills in police officers. *Applied Cognitive Psychology*, *33*(5), 828–842. https://doi.org/10.1002/acp.3525

Bate, S., Mestry, N., & Portch, E. (2021). Individual differences between observers in face matching. In M. Bindemann (Ed.), *Forensic Face Matching*. Oxford University Press.

Bate, S., & Murray, E. (2017). Extremes of facial recognition: prosopagnosia and super recognition. In M. Bindemann & A. M. Megreya (Eds.), *Face Processing: Systems, Disorders and Cultural Difficulties* (pp. 203–222). Nova Science Publishers.

Beveridge, J. R., Phillips, P. J., Givens, G. H., Draper, B. A., Teli, M. N., & Bolme, D. S. (2011). When high-quality face images match poorly. *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*, 572–578. https://doi.org/10.1109/FG.2011.5771460

Bindemann, M., Attard, J., & Johnston, R. A. (2014). Perceived ability and actual recognition accuracy for unfamiliar and famous faces. *Cogent Psychology*, *1*(1). https://doi.org/10.1080/23311908.2014.986903

Bindemann, M., Attard, J., Leach, A., & Johnston, R. A. (2013). The effect of image pixelation on unfamiliar-face matching. *Applied Cognitive Psychology*, *27*(6), 707–717. https://doi.org/10.1002/acp.2970

Bindemann, M., Avetisyan, M., & Rakow, T. (2012). Who can recognize unfamiliar faces? Individual differences and observer consistency in person identification. *Journal of Experimental Psychology. Applied*, *18*(3), 277–291. https://doi.org/10.1037/a0029635

Bindemann, M., & Burton, A. M. (2021). Steps towards a cognitive theory of facial identity comparison. In M. Bindemann (Ed.), *Forensic Face Matching*. Oxford University Press.

Bindemann, M., Fysh, M., Cross, K., & Watts, R. (2016). Matching faces against the clock. *I-Perception*, *7*(5), 1–18. https://doi.org/10.1177/2041669516672219

Bindemann, M., & Sandford, A. (2011). Me, myself, and I: Different recognition rates for three photo-IDs of the same person. *Perception*, *40*(5), 625–627. https://doi.org/10.1068/p7008

Bobak, A. K., Bennetts, R. J., Parris, B. A., Jansari, A., & Bate, S. (2016). An in-depth cognitive examination of individuals with superior face recognition skills. *Cortex*, *82*, 48–62. https://doi.org/10.1016/j.cortex.2016.05.003

Bobak, A. K., Dowsett, A. J., & Bate, S. (2016). Solving the border control problem: Evidence of enhanced face matching in individuals with extraordinary face recognition skills. *PLoS ONE*, *11*(2). https://doi.org/10.1371/journal.pone.0148148

Bobak, A. K., Hancock, P. J. B., & Bate, S. (2016). Super-recognisers in Action: Evidence from Face-matching and Face Memory Tasks. *Applied Cognitive Psychology*, *30*(1), 81–91. https://doi.org/10.1002/acp.3170

Bobak, A. K., Pampoulov, P., & Bate, S. (2016). Detecting superior face recognition skills in a large sample of young British adults. *Frontiers in Psychology*, *7*(SEP). https://doi.org/10.3389/fpsyg.2016.01378

Bobak, A. K., Parris, B. A., Gregory, N. J., Bennetts, R. J., & Bate, S. (2017). Eye-movement strategies in developmental prosopagnosia and "super" face recognition. *The Quarterly Journal of Experimental Psychology*, *70*(2), 201–217. https://doi.org/10.1080/17470218.2016.1161059

Bollé, T., Casey, E., & Jacquet, M. (2020). The role of evaluations in reaching decisions using automated systems supporting forensic analysis. *Forensic Science International: Digital Investigation*, *34*, 301016. https://doi.org/10.1016/j.fsidi.2020.301016

Booth, R. (2020, March 12). Halt public use of facial recognition tech, says equality watchdog. *The Guardian*. %0A This article is more than 2 months old%0AHalt public use of facial recognition tech, says equality watchdog

Bromby, M. (2006). CCTV and Expert Evidence: Addressing the Reliability of New Sciences. *Archbold News*, *9*, 6–9.

Bromby, M., & Plews, S. (2003). Facing up to Change? *E-Law Review*, *13*.

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology*, *73*, 105–116.

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, *7*(3), 207–218. https://doi.org/10.1037/1076-898X.7.3.207

Bruce, V., & Young, A. W. (1986). Understanding face recognition. *British Journal of Psychology*, *77*, 305–327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Bruveris, M., Gietema, J., Mortazavian, P., & Mahadevan, M. (2020). Reducing Geographic Performance Differential for Face Recognition. *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 98–106. 10.1109/WACVW50321.2020.9096930

Buchanan, B., Feigenbaum, E., & Lederberg, J. (1969). Heuristic DENDRAL - A program for generating explanatory hypotheses in organic chemistry. *Machine Intelligence 4*, *4*(February 1968), 209–254.

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *Quarterly Journal of Experimental Psychology*, *66*(8), 1467–1485. https://doi.org/10.1080/17470218.2013.800125

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity From Variation: Representations of Faces Derived From Multiple Instances. *Cognitive*

*Science*, *40*(1), 202–223. https://doi.org/10.1111/cogs.12231

Burton, A. M., Miller, P., Bruce, V., Hancock, P. J. B., & Henderson, Z. (2001). Human and automatic face recognition: A comparison across image formats. *Vision Research*, *41*(24), 3185–3195. https://doi.org/10.1016/S0042-6989(01)00186-9

Burton, A. M., Schweinberger, S. R., Jenkins, R., & Kaufmann, J. M. (2015). Arguments Against a Configural Processing Account of Familiar Face Recognition. *Perspectives on Psychological Science*, *10*(4), 482–496. https://doi.org/10.1177/1745691615583129

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, *42*(1), 286–291. https://doi.org/10.3758/BRM.42.1.286

Burton, M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face Recognition in Poor-Quality Video: Evidence from Security Surveillance. *Psychological Science*, *10*(3), 243–248. https://doi.org/10.1111/1467-9280.00144

Cantor, M., Aplin, L. M., & Farine, D. R. (2020). A primer on the relationship between group size and group performance. *Animal Behaviour*, *166*, 139–146. https://doi.org/10.1016/j.anbehav.2020.06.017

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). VGGFace2: A dataset for recognising faces across pose and age. *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, 67–74. https://doi.org/10.1109/FG.2018.00020

Caplova, Z., Obertova, Z., Gibelli, D. M., Mazzarelli, D., Fracasso, T., Vanezis, P., Sforza, C., & Cattaneo, C. (2017). The Reliability of Facial Recognition of Deceased Persons on Photographs. *Journal of Forensic Sciences*, *6*, 1–6. https://doi.org/10.1111/1556-4029.13396

Carmel, D., & Bentin, S. (2002). Domain specificity versus expertise: Factors influencing distinct processing of faces. *Cognition*, *83*(1), 1–29. https://doi.org/10.1016/S0010-0277(01)00162-7

Choi, H., & Watanabe, T. (2012). Perceptual learning solely induced by feedback. *Vision Research*, *61*, 77–82. https://doi.org/10.1016/j.visres.2012.01.006

Costen, N. P., Parker, D. M., & Craw, I. (1996). Effects of high-pass and low-pass spatial filtering on face identification. *Perception & Psychophysics*, *58*(4), 602–612. https://doi.org/10.3758/BF03213093

Crawford, J. R., Garthwaite, P. H., & Porter, S. (2010). Point and interval estimates of effect sizes for the case-controls design in neuropsychology: Rationale, methods, implementations, and proposed reporting standards. *Cognitive Neuropsychology*, *27*(3), 245–260. https://doi.org/10.1080/02643294.2010.513967

Crawford, J. R., Garthwaite, P. H., & Ryan, K. (2011). Comparing a single case to a control sample: Testing for neuropsychological deficits and dissociations in the presence of covariates. *Cortex*, *47*(10), 1166–1178. https://doi.org/10.1016/j.cortex.2011.02.017

Davis, J. P. (2019). The worldwide public impact of identifying super-recognisers for police and business. *Cognitive Psychology Bulletin*, *4*(July), 17–21.

Davis, J. P., Lander, K., Evans, R., & Jansari, A. (2016). Investigating Predictors of Superior Face Recognition Ability in Police Super-recognisers. *Applied Cognitive Psychology*. https://doi.org/10.1002/acp.3260

Davis, J. P., Maigut, A., & Forrest, C. (2019). The wisdom of the crowd: A case of post- to ante-mortem face matching by police super-recognisers. *Forensic Science International*, *302*, 109910. https://doi.org/10.1016/j.forsciint.2019.109910

Davis, J. P., & Robertson, D. J. (2020). Capitalizing on the Super- Recognition Advantage. *Journal of the Homeland Defense and Security Information Analysis Center*, *Spring*(June), 20–25. https://www.hdiac.org

Devue, C. (2019). Breaking face processing tasks apart to improve their predictive value in the real world: A comment on Ramon, Bobak, and White (2019). *British Journal of Psychology*, *110*(3), 483–485. https://doi.org/10.1111/bjop.12391

Diamond, R., & Carey, S. (1986). Why Faces Are and Are Not Special. An Effect of Expertise. *Journal of Experimental Psychology: General*, *115*(2), 107–117. https://doi.org/10.1037/0096-3445.115.2.107

Dowsett, A. J., & Burton, A. M. (2015). Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *British Journal of Psychology*, *106*(3), 433–445. https://doi.org/10.1111/bjop.12103

Dror, I. E. (2011). Why Experts Get It Wrong. In N. Kapur (Ed.), *The Paradoxical Brain* (pp. 177–188). Cambridge University Press.

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*(4), 576–585. https://doi.org/10.1016/j.neuropsychologia.2005.07.001

Dunn, J. D., Summersby, S., Towler, A., Davis, J., & White, D. (2020). *UNSW Face Test: A screening tool for super-recognizers*. 1–19. https://doi.org/10.31234/osf.io/k7mf6

Durkin, J. (1990). Research Review: Application of expert systems in the sciences. *The Ohio Journal of Science*, *90*(5), 171–179.

Edmond, G., Biber, K., Kemp, R., & Porter, G. (2009). Law's Looking Glass: Expert Identification Evidence Derived from Photographic and Video Images. *Current Issues in Criminal Justice*, *20*(3), 337–377.

Edmond, G., Towler, A., Growns, B., Ribeiro, G., Found, B., White, D., Ballantyne, K., Searston, R. A., Thompson, M. B., Tangen, J. M., Kemp, R. I., & Martire, K. (2017). Thinking forensics: Cognitive science for forensic practitioners. *Science & Justice*, *57*(2), 144–154. https://doi.org/10.1016/j.scijus.2016.11.005

Engstrom, T. E. J. (2003). Sharing knowledge through mentoring. *Performance Improvement*, *42*(8), 36–42.

Ericsson, K. A., & Lehmann, A. C. (1996). EXPERT AND EXCEPTIONAL PERFORMANCE: Evidence of Maximal Adaptation to Task Constraints. *Annual Review of Psychology*, *47*(1), 273–305. https://doi.org/10.1146/annurev.psych.47.1.273

Ericsson, K. A., & Staszewski, J. J. (1989). Skilled memory and expertise: Mechanisms of exceptional performance. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon* (pp. 235–267). Psychology Press.

Etchells, D., Brooks, J., & Johnston, R. (2016). Evidence for view-invariant Face Recognition Units in unfamiliar face learning. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1080/00335558008248231

European Network of Forensic Science Institutes. (2015). *ENFSI Best Practice Manual for Fingerprint Examination* (Vol. 01, Issue November). http://enfsi.eu/wp-content/uploads/2016/09/6._fingerprint_examination_0.pdf

European Network of Forensic Science Institutes. (2018). *ENFSI Best Practice Manual for Facial Image Comparison* (Vol. 01, Issue January). http://enfsi.eu/wp-content/uploads/2017/06/ENFSI-BPM-DI-01.pdf

Facial Identification Scientific Working Group. (2010). *Guidelines and Recommendations for Facial Comparison Training to Competency*. Facial Identification Scientific Working Group. https://fiswg.org/FISWG_Training_Guidelines_Recommendations_v1.1_2010_11_18.pdf

Facial Identification Scientific Working Group. (2012b). *Recommendations for a Training Program in Facial Comparison*. https://fiswg.org/FISWG_RecommendationsForTrainingProgram_v1.0_2012_02_02.pdf

Facial Identification Scientific Working Group. (2018). *Facial Image Comparison Feature List for Morphological Analysis*. https://fiswg.org/FISWG_Morph_Analysis_Feature_List_v2.0_20180911.pdf

Facial Identification Scientific Working Group. (2019a). *Facial Comparison Overview and Methodology Guidelines*. https://fiswg.org/fiswg_facial_comparison_overview_and_methodology_guidelines_V1.0_20191025.pdf

Facial Identification Scientific Working Group. (2019b). *Guide for Facial Comparison Training of Examiners to Competency*.

Facial Identification Scientific Working Group. (2019c). *Guide for Facial Comparison Training of Reviewers to Competency*.

Facial Identification Scientific Working Group. (2019d). *Guide for Mentorship of Facial Comparison Trainees in Role Based Facial Comparison*. https://fiswg.org/FISWG_Mentorship_for_Facial_Comparison_Trainees_v1.0_20190510.pdf

Feng, X., & Burton, A. M. (2019). Identity Documents Bias Face Matching. *Perception*, *48*(12), 1163–1174. https://doi.org/10.1177/0301006619877821

FISWG. (2019). *Guide for mentorship of facial comparison trainees in role based facial comparison, version 1.0*. https://fiswg.org/FISWG_Mentorship_for_Facial_Comparison_Trainees_v1.0_20190510.pdf

FISWG. (2020). *Minimum Training Criteria for Assessors Using Facial Recognition Systems*.

Fong, R. C., & Vedaldi, A. (2017). Interpretable Explanations of Black Boxes by Meaningful Perturbation. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*, 3449–3457. https://doi.org/10.1109/ICCV.2017.371

Forensic Science Regulator. (2018). *Annual Report November 2016 - November 2017*.

Fysh, M. C., & Bindemann, M. (2017a). Forensic Face Matching: A Review. In M. Bindemann & A. M. Megreya (Eds.), *Face Processing: Systems, Disorders and Cultural Difficulties* (pp. 1–20). Nova Science Publishers.

Fysh, M. C., & Bindemann, M. (2017b). The Kent Face Matching Test. *British Journal of Psychology*, 1–13. https://doi.org/10.1111/bjop.12260

Fysh, M. C., & Bindemann, M. (2017c). Effects of time pressure and time passage on face-matching accuracy. *Royal Society Open Science*, *4*(6), 1–13. https://doi.org/10.1098/rsos.170249

Fysh, M. C., & Bindemann, M. (2018). Human–Computer Interaction in Face Matching. *Cognitive Science*, *42*(5), 1714–1732. https://doi.org/10.1111/cogs.12633

Fysh, M. C., Stacchi, L., & Ramon, M. (2020). Differences between and within individuals, and subprocesses of face cognition: implications for theory, research and personnel selection. *Royal Society Open Science*, *7*(9), 200233. https://doi.org/10.1098/rsos.200233

Gabrielson, R. (2019, January 17). The FBI Says Its Photo Analysis Is Scientific Evidence. Scientists Disagree. *Propublica*. https://www.propublica.org/article/with-photo-analysis-fbi-lab-continues-shaky-forensic-science-practices

Gauthier, I., Anderson, A. W., Skudlarski, P., & Gore, J. C. (2000). Expertise for cars and birds recruits right hemisphere face areas. *Nature Neuroscience*, 19C.

Gentry, N. W., & Bindemann, M. (2019). Examples Improve Facial Identity Comparison. *Journal of Applied Research in Memory and Cognition*, *8*(3), 376–385. https://doi.org/10.1016/j.jarmac.2019.06.002

Ghuman, A. S., Brunet, N. M., Li, Y., Konecky, R. O., Pyles, J. A., Walls, S. A., Destefino, V., Wang, W., & Richardson, R. M. (2014). Dynamic encoding of face information in the human fusiform gyrus. *Nature Communications*, *5*. https://doi.org/10.1038/ncomms6672

Gobet, F., & Charness, N. (2006). Expertise in Chess. In K. Anders Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (pp. 523–538). Cambridge University Press. https://doi.org/10.1017/CBO9780511816796.030

Grother, P., Ngan, M., & Hanaoka, K. (2019a). Face Recognition Vendor Test (FRVT) Part 2: Identification. In *NISTIR 8271*. https://doi.org/https://doi.org/10.6028/NIST.IR.8271

Grother, P., Ngan, M., & Hanaoka, K. (2019b). Face Recognition Vendor Test ( FRVT ) Part 3 : Demographic Effects. In *NISTIR 8280*. https://doi.org/https://doi.org/10.6028/NIST.IR.8280

Growns, B., & Martire, K. A. (2020a). Forensic Feature-Comparison Expertise: Statistical Learning Facilitates Visual Comparison Performance. *Journal of Experimental Psychology: Applied*. https://doi.org/10.1037/xap0000266

Growns, B., & Martire, K. A. (2020b). Human factors in forensic science: The cognitive mechanisms that underlie forensic feature-comparison expertise. *Forensic Science International: Synergy*, *2*, 148–153. https://doi.org/10.1016/j.fsisyn.2020.05.001

Hackman, J. R., & Morris, C. G. (1975). Group tasks, group interaction process, and group performance effectiveness: A review and proposed integration. *Advances in Experimental Social Psychology*, *8*(C), 45–99. https://doi.org/10.1016/S0065-2601(08)60248-8

Hancock, P. J. B., Somai, R. S., & Mileva, V. R. (2020). Convolutional neural net face recognition works in non-human-like ways. *Royal Society Open Science*, *7*(10), 1–7. https://doi.org/10.1098/rsos.200595rsos200595

Heyer, R. (2013). *Understanding One-to-Many Unfamiliar Face Matching in the Operational Context : The Impact of Candidate List Size , Expertise , and Decision Aids on the Performance of Facial Recognition System Users* (Issue October). University of Adelaide.

Heyer, R., Semmler, C., & Hendrickson, A. T. (2018). Humans and Algorithms for Facial Recognition: The Effects of Candidate List Length and Experience on Performance. *Journal of Applied Research in Memory and Cognition*, *7*(4), 597–609. https://doi.org/10.1016/j.jarmac.2018.06.002

Hill, K. (2020, June 4). Wrongfully Accused by an Algorithm. *New York Times*. 2020

Hill, M. Q., Parde, C. J., Castillo, C. D., Colón, Y. I., Ranjan, R., Chen, J.-C., Blanz, V., & O'Toole, A. J. (2019). Deep convolutional neural networks in the face of caricature. *Nature Machine Intelligence*, *1*(11), 522–529. https://doi.org/10.1038/s42256-019-0111-7

Hole, G. J., George, P. A., Eaves, K., & Rasek, A. (2002). Effects of geometric distortions on face-recognition performance. *Perception*, *31*(10), 1221–1240. https://doi.org/10.1068/p3252

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups

of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(46), 16385–16389. https://doi.org/10.1073/pnas.0403723101

Houlton, T. M. R., & Steyn, M. (2018). Finding Makhubu: A morphological forensic facial comparison. *Forensic Science International*, *285*(February), 13–20. https://doi.org/10.1016/j.forsciint.2018.01.022

Howard, J. J., Rabbitt, L. R., & Sirotin, Y. B. (2020). Human-algorithm teaming in face recognition: How algorithm outcomes cognitively bias human decision-making. *PLoS ONE*, *15*(8 August 2020), 1–18. https://doi.org/10.1371/journal.pone.0237855

Hussain, Z., Bennett, P. J., & Sekuler, A. B. (2012). Versatile perceptual learning of textures after variable exposures. *Vision Research*, *61*, 89–94. https://doi.org/10.1016/j.visres.2012.01.005

Hussain, Z., McGraw, P. V., Sekuler, A. B., & Bennett, P. J. (2012). The rapid emergence of stimulus specific perceptual learning. *Frontiers in Psychology*, *3*(JUL), 1–5. https://doi.org/10.3389/fpsyg.2012.00226

Hussain, Z., Sekuler, A. B., & Bennett, P. J. (2009a). How much practice is needed to produce perceptual learning? *Vision Research*, *49*(21), 2624–2634. https://doi.org/10.1016/j.visres.2009.08.022

Hussain, Z., Sekuler, A. B., & Bennett, P. J. (2009b). Perceptual learning modifies inversion effects for faces and textures. *Vision Research*, *49*(18), 2273–2284. https://doi.org/10.1016/j.visres.2009.06.014

Iida, R., Itsukusima, Y., & Mah, E. Y. (2020). How do we judge our confidence? Differential effects of meta-memory feedback on eyewitness accuracy and confidence. *Applied Cognitive Psychology*, *34*(2), 397–408. https://doi.org/10.1002/acp.3625

Ingram, M., Moreton, R., Gancz, B., & Pollick, F. (*in press*.). Calibrating Trust Towards an Autonomous Image Classifier. *Technology, Mind, and Behavior*.

Jacquet, M., & Champod, C. (2020). Automated face recognition in forensic science: Review and perspectives. *Forensic Science International*, *307*. https://doi.org/10.1016/j.forsciint.2019.110124

Jeckeln, G., Hahn, C. A., Noyes, E., Cavazos, J. G., & O'Toole, A. J. (2018). Wisdom of the social versus non-social crowd in face identification. *British Journal of Psychology*, 1–12. https://doi.org/10.1111/bjop.12291

Jenkins, R., White, D., Van Montfort, X., & Mike Burton, A. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313–323. https://doi.org/10.1016/j.cognition.2011.08.001

Kahneman, D. (2011). *Thinking, Fast and Slow*. Penguin.

Kassin, S. M., Dror, I. E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and*

*Cognition*, *2*(1), 42–52. https://doi.org/10.1016/j.jarmac.2013.01.001

Kemp, R., Towell, N., & Pike, G. (1997). When Seeing should not be Believing: Photographs, Credit Cards and Fraud. *Applied Cognitive Psychology*, *11*(3), 211–222. https://doi.org/10.1002/(SICI)1099-0720(199706)11:3<211::AID-ACP430>3.0.CO;2-O

Keval, H. U., & Sasse, M. A. (2008). Can we ID from CCTV: Image quality in digital CCTV and face identification performance. *Proceedings SPIE*, *6982*, 69820K. https://doi.org/10.1117/12.774212

Kleinberg, K. F., & Vanezis, P. (2007). Variation in proportion indices and angles between selected facial landmarks with rotation in the Frankfort plane. *Medicine, Science, and the Law*, *47*(2), 107–116. https://doi.org/10.1258/rsmmsl.47.2.107

Kleinberg, K. F., Vanezis, P., & Burton, A. M. (2007). Failure of anthropometry as a facial identification technique using high-quality photographs. *Journal of Forensic Sciences*, *52*(4), 779–783. https://doi.org/10.1111/j.1556-4029.2007.00458.x

Kramer, R. S. S., & Ritchie, K. L. (2016). Disguising Superman: How Glasses Affect Unfamiliar Face Matching. *Applied Cognitive Psychology*, *30*(6), 841–845. https://doi.org/10.1002/acp.3261

Kramer, R. S. S., Young, A. W., & Burton, A. M. (2018). Understanding face familiarity. *Cognition*, *172*(June 2017), 46–58. https://doi.org/10.1016/j.cognition.2017.12.005

Krause, S., James, R., Faria, J. J., Ruxton, G. D., & Krause, J. (2011). Swarm intelligence in humans: Diversity can trump ability. *Animal Behaviour*, *81*(5), 941–948. https://doi.org/10.1016/j.anbehav.2010.12.018

Lander, K., Bruce, V., & Bindemann, M. (2018). Use-inspired basic research on individual differences in face identification: implications for criminal investigation and security. *Cognitive Research: Principles and Implications*, *3*(1), 26. https://doi.org/10.1186/s41235-018-0115-6

Langdon, F. J. (2014). Evidence of mentor learning and development: an analysis of New Zealand mentor/mentee professional conversations. *Professional Development in Education*, *40*(1), 36–55. https://doi.org/10.1080/19415257.2013.833131

Layher, E., Dixit, A., & Miller, M. B. (2020). Who gives a criterion shift? A uniquely individualistic cognitive trait. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *46*(11), 2075–2105. https://doi.org/10.1037/xlm0000951

Lee, W.-L., Wilkinson, C., Memon, A., & Houston, K. (2006). Matching unfamiliar faces from poor quality closed-circuit television (CCTV) footage: An evaluation of the effect of training on facial identification ability. *AXIS Online Jounal of Centre for Anatomy and Human Identification*, *1*(1), 19–28.

Lucas, H. D., Chiao, J. Y., & Paller, K. A. (2011). Why some faces won't be remembered: Brain potentials illuminate successful versus unsuccessful encoding for same-race and other-race faces. *Frontiers in Human Neuroscience*, *5*(MARCH), 1–17. https://doi.org/10.3389/fnhum.2011.00020

Macarulla Rodriguez, A., Geradts, Z., & Worring, M. (2020). Likelihood Ratios for Deep Neural Networks in Face Comparison. *Journal of Forensic Sciences*, *65*(4), 1169–1183. https://doi.org/10.1111/1556-4029.14324

Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S. J., & Frith, C. D. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(8), 4398–4403. https://doi.org/10.1073/pnas.070039597

Mallett, X., & Evison, M. P. (2013). Forensic facial comparison: Issues of admissibility in the development of novel analytical technique. *Journal of Forensic Sciences*, *58*(4), 859–865. https://doi.org/10.1111/1556-4029.12127

Mangiafico, S. (2019). *rcompanion: Functions to Support Extension Education Program Evaluation* (R package version 2.2.1).

Martire, K. A., & Kemp, R. I. (2016). Considerations when designing human performance tests in the forensic sciences. *Australian Journal of Forensic Sciences*, *0618*(November), 1–17. https://doi.org/10.1080/00450618.2016.1229815

Masi, I., Wu, Y., Hassner, T., & Natarajan, P. (2019). Deep Face Recognition: A Survey. *Proceedings - 31st Conference on Graphics, Patterns and Images, SIBGRAPI 2018*, 471–478. https://doi.org/10.1109/SIBGRAPI.2018.00067

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, *6*(6), 255–260. https://doi.org/10.1016/S1364-6613(02)01903-4

Mayfield, M. (2010). Tacit knowledge sharing: Techniques for putting a powerful tool in practice. *Development and Learning in Organisations*, *24*(1), 24–26. https://doi.org/10.1108/14777281011010497

McKone, E., Kanwisher, N., & Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? *Trends in Cognitive Sciences*, *11*(1), 8–15. https://doi.org/10.1016/j.tics.2006.11.002

McNeill, M. A., Suchomska, M., & Strathie, A. (2015). Expert facial comparison evidence: Science versus pseudo science. *Psychology and Law*, *5*(4), 127–140. https://doi.org/10.17759/psylaw.2015050411

Medsker, L., & Turban, E. (1994). Integrating expert systems and neural computing for decision support. *Expert Systems with Applications*, *7*(4), 553–562.

Megreya, A. M. (2018). Feature-by-feature comparison and holistic processing in unfamiliar face matching. *PeerJ*, *6*, e4437. https://doi.org/10.7717/peerj.4437

Megreya, A. M., & Bindemann, M. (2018). Feature instructions improve face-matching accuracy. *PLoS ONE*, *13*(3), 1–16. https://doi.org/10.1371/journal.pone.0193455

Megreya, A. M., & Burton, a M. (2006). Unfamiliar faces are not faces: evidence from a matching task. *Memory & Cognition*, *34*(4), 865–876. https://doi.org/10.3758/BF03193433

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception and Psychophysics*, *69*(7), 1175–1184. https://doi.org/10.3758/BF03193954

Megreya, A. M., Sandford, A., & Burton, A. M. (2013). Matching face images raken on the same day or months apart: The limitations of photo ID. *Applied Cognitive Psychology*, *27*(6), 700–706. https://doi.org/10.1002/acp.2965

Megreya, A. M., White, D., & Burton, A. M. (2011). The other-race effect does not rely on memory: Evidence from a matching task. *The Quarterly Journal of Experimental Psychology*, *64*(8), 1473–1483. https://doi.org/10.1080/17470218.2011.575228

Meissner, C. A., & Brigham, J. C. (2001). Thirty Years of Investigating the Own-Race Bias in Memory for Faces: A Meta-Analytic Review. *Psychology, Public Policy, and Law*, *7*(1), 3–35. https://doi.org/10.1037/1076-8971.7.1.3

Mileva, M., & Burton, A. M. (2019). Face search in CCTV surveillance. *Cognitive Research: Principles and Implications*, *4*(1). https://doi.org/10.1186/s41235-019-0193-0

Moreton, R. (2021). Forensic face matching: Procedures and application. In M. Bindemann (Ed.), *Forensic Face Matching*. Oxford University Press.

Moreton, R., & Morley, J. (2011). Investigation into the use of photoanthropometry in facial image comparison. *Forensic Science International*, *212*(1–3), 231–237. https://doi.org/10.1016/j.forsciint.2011.06.023

Moreton, R., Pike, G., & Havard, C. (2019). A task- and role-based perspective on super-recognizers: Commentary on 'Super-recognizers: From the laboratory to the world and back again.' *British Journal of Psychology*, bjop.12394. https://doi.org/10.1111/bjop.12394

Moshakis, A. (2018, November 11). Super recognisers: the people who never forget a face. *The Guardian*. https://www.theguardian.com/uk-news/2018/nov/11/super-recognisers-police-the-people-who-never-forget-a-face

National Academy of Science. (2009). Strengthening Forensic Science in the United States: A Path Forward. In *National Academies Press*. https://doi.org/10.1016/0379-0738(86)90074-5

Norell, K., Läthén, K. B., Bergström, P., Rice, A., Natu, V., & O'Toole, A. (2015). The Effect of Image Quality and Forensic Expertise in Facial Image Comparisons. *Journal of Forensic Sciences*, *60*(2), 331–340. https://doi.org/10.1111/1556-4029.12660

Noyes, E. (2016). *Face Recognition in Challenging Situations* (Issue May).

Noyes, E., & Hill, M. Q. (2021). Automatic recognition systems and human computer interaction in face matching. In M. Bindemann (Ed.), *Forensic Face Matching*. Oxford University Press.

Noyes, E., & O'Toole, A. J. (2017). Face recognition assessments used in the study of super-recognisers. *ArXiv*, *June*, arXiv1705.04739.

Noyes, E., Phillips, P. J., & O'Toole, A. J. (2017). What is a Super-Recogniser. In M. Bindemann & A. M. Megreya (Eds.), *Face Processing: Systems, Disorders and Cultural Difficulties* (pp. 173–202). Nova Science Publishers.

O'Toole, A. J., Abdi, H., Jiang, F., & Phillips, P. J. (2007). Fusing face-verification algorithms and humans. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *37*(5), 1149–1155. https://doi.org/10.1109/TSMCB.2007.907034

O'Toole, A. J., An, X., Dunlop, J., Natu, V., & Phillips, P. J. (2012). Comparing face recognition algorithms to humans on challenging tasks. *ACM Transactions on Applied Perception*, *9*(4), 1–15. https://doi.org/10.1145/2355598.2355599

O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q., & Chellappa, R. (2018). Face Space Representations in Deep Convolutional Neural Networks. *Trends in Cognitive Sciences*, *22*(9), 794–809. https://doi.org/10.1016/j.tics.2018.06.006

O'Toole, A. J., Edelman, S., & Bülthoff, H. H. (1998). Stimulus-specific effects in face recognition over changes in viewpoint. *Vision Research*, *38*(15–16), 2351–2363. https://doi.org/10.1016/S0042-6989(98)00042-X

O'Toole, A. J., Phillips, P. J., & Narvekar, A. (2008). Humans versus algorithms: Comparisons from the face recognition vendor test 2006. *2008 8th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2008*, *January 2015*. https://doi.org/10.1109/AFGR.2008.4813318

Papesh, M. H., Heisick, L. L., & Warner, K. A. (2018). The persistent low-prevalence effect in unfamiliar face-matching: The roles of feedback and criterion shifting. *Journal of Experimental Psychology: Applied*, *24*(3), 416–430. https://doi.org/10.1037/xap0000156

Phillips, P. J., Moon, H., Rauss, P., & Rizvi, S. A. (1997). FERET evaluation methodology for face-recognition algorithms. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 137–143. https://doi.org/10.1109/cvpr.1997.609311

Phillips, P. J., & O'Toole, A. J. (2014). Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing*, *32*(1), 74–85. https://doi.org/10.1016/j.imavis.2013.12.002

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., Chen, J.-C., Castillo, C. D., Chellappa, R., White, D., & O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 201721355. https://doi.org/10.1073/pnas.1721355115

President's Committee of Advisors on Science and Technology. (2016). *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. *September*, 1–174.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

R v Stockwell, (1993). http://login.westlaw.co.uk/maf/wluk/ext/app/document?docguid=I6AF24CE0E42811DA8FC2A0F0355337E9&crumb-action=reset&entityID=https://idp.dundee.ac.uk/shibboleth

Ramon, M., Bobak, A. K., & White, D. (2019a). Super-recognizers: From the lab to the world and back again. *British Journal of Psychology*, *110*(3), 461–479. https://doi.org/10.1111/bjop.12368

Ramon, M., Bobak, A. K., & White, D. (2019b). Towards a 'manifesto' for super-recognizer research. *British Journal of Psychology*, *110*(3), 495–498. https://doi.org/10.1111/bjop.12411

Ritz-Timme, S., Gabriel, P., Obertovà, Z., Boguslawski, M., Mayer, F., Drabik, A., Poppa, P., De Angelis, D., Ciaffi, R., Zanotti, B., Gibelli, D., & Cattaneo, C. (2011). A new atlas for the evaluation of facial features: Advantages, limits, and applicability. *International Journal of Legal Medicine*, *125*(2), 301–306. https://doi.org/10.1007/s00414-010-0446-4

Robertson, D. J., & Burton, A. M. (2021). Checking ID-cards for the sale of restricted goods: Age decisions bias face decisions. *Applied Cognitive Psychology*, *35*(1), 71–81. https://doi.org/10.1002/acp.3739

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by metropolitan police super-recognisers. *PLoS ONE*, *11*(2), 1–8. https://doi.org/10.1371/journal.pone.0150036

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Robson, S. G., Searston, R. A., Edmond, G., McCarthy, D. J., & Tangen, J. M. (2020). An expert–novice comparison of feature choice. *Applied Cognitive Psychology*, *April*, 1–12. https://doi.org/10.1002/acp.3676

Rosenberg, L. (2016). Artificial Swarm Intelligence vs human experts. *Proceedings of the International Joint Conference on Neural Networks*, *2016-Octob*, 2547–2551. https://doi.org/10.1109/IJCNN.2016.7727517

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: people with extraordinary face recognition ability. *Psychonomic Bulletin & Review*, *16*(2), 252–257. https://doi.org/10.3758/PBR.16.2.252

Sanchez, C., & Dunning, D. (2018). Overconfidence among beginners: Is a little learning a dangerous thing? *Journal of Personality and Social Psychology*, *114*(1), 10–28. https://doi.org/10.1037/pspa0000102

Sauerland, M., Sagana, A., Siegmann, K., Heiligers, D., Merckelbach, H., & Jenkins, R. (2016). These two are different. Yes, they're the same: Choice blindness for facial identity. *Consciousness and Cognition*, *40*, 93–104. https://doi.org/10.1016/j.concog.2016.01.003

Searston, R. A., & Tangen, J. M. (2017a). Training perceptual experts: Feedback, labels, and contrasts. *Canadian Journal of Experimental Psychology*, *71*(1), 32–39. https://doi.org/10.1037/cep0000124

Searston, R. A., & Tangen, J. M. (2017b). The Emergence of Perceptual Expertise with Fingerprints Over Time. *Journal of Applied Research in Memory and Cognition*, *6*(4), 442–451. https://doi.org/10.1016/j.jarmac.2017.08.006

Seckiner, D., Mallett, X., Roux, C., Meuwly, D., & Maynard, P. (2018). Forensic image analysis – CCTV distortion and artefacts. *Forensic Science International*, *285*, 77–85. https://doi.org/10.1016/j.forsciint.2018.01.024

Seitz, A. R., & Dinse, H. R. (2007). A common framework for perceptual learning. *Current Opinion in Neurobiology*, *17*(2), 148–153. https://doi.org/10.1016/j.conb.2007.02.004

Sita, J., Found, B., & Rogers, D. K. (2002). Forensic Handwriting Examiners' Expertise for Signature Comparison. *Journal of Forensic Sciences*, *47*(5), 15521J. https://doi.org/10.1520/jfs15521j

Skovholt, T. M., Hanson, M., Jennings, L., & Grier, T. (2016). A Brief History of Expertise. In T. M. Skovholt & L. Jennings (Eds.), *Master Therapists: Exploring Expertise in Therapy and Counseling* (10th ed., Issue August, pp. 1–7). Oxford University Press. https://doi.org/10.1093/med:psych/9780190496586.001.0001

Sosik, J. J., Godshalk, V. M., & Yammarino, F. J. (2004). Transformational leadership, learning goal orientation, and expectations for career success in mentor-protégé relationships: A multiple levels of analysis perspective. *Leadership Quarterly*, *15*(2), 241–261. https://doi.org/10.1016/j.leaqua.2004.02.003

Spaun, N. A. (2009). Facial comparisons by subject matter experts: Their role in biometrics and their training. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5558 LNCS*, 161–168. https://doi.org/10.1007/978-3-642-01793-3_17

Stacchi, L., Huguenin-Elie, E., Caldara, R., & Ramon, M. (2019). Normative data for two ecologically valid tests of face identity matching. *Journal of Vision*, *19*(10), 136a. https://doi.org/10.1167/19.10.136a

Stephens, R. G., Semmler, C., & Sauer, J. D. (2017). The effect of the proportion of mismatching trials and task orientation on the confidence-accuracy relationship in unfamiliar face matching. *Journal of Experimental Psychology: Applied*, *23*(3), 336–353. https://doi.org/10.1037/xap0000130

Stevens, C. (2021). Person identification at airports during passport control. In M. Bindemann (Ed.), *Forensic Face Matching*. Oxford University Press.

Steyn, M., Pretorius, M., Briers, N., Bacci, N., Johnson, A., & Houlton, T. M. R. (2018). Forensic facial comparison in South Africa: State of the science. *Forensic Science International*, *287*, 190–194. https://doi.org/10.1016/J.FORSCIINT.2018.04.006

Strathie, A., & McNeill, A. (2016). Facial Wipes don't Wash: Facial Image Comparison by

Video Superimposition Reduces the Accuracy of Face Matching Decisions. *Applied Cognitive Psychology*, *30*(4), 504–513. https://doi.org/10.1002/acp.3218

Strathie, A., Mcneill, A., & White, D. (2012). In the Dock: Chimeric Image Composites Reduce Identification Accuracy. *Applied Cognitive Psychology*, *26*(1), 140–148. https://doi.org/10.1002/acp.1806

Surowiecki, J. (2004). *The Wisdom of Crowds*. Abacus.

Swap, W., Leonard, D., Shields, M., & Abrams, L. (2001). Using Mentoring and Storytelling to Transfer Knowledge in the Workplace. *Journal of Management Information Systems*, *18*(1), 95–114. https://doi.org/10.1080/07421222.2001.11045668

Swofford, H., & Champod, C. (2021). Implementation of Algorithms in Pattern & Impression Evidence: A Responsible and Practical Roadmap. *Forensic Science International: Synergy*. https://doi.org/https://doi.org/10.1016/j.fsisyn.2021.100142

Tanaka, J. W., Curran, T., & Sheinberg, D. L. (2005). The training and transfer of real-world perceptual expertise. *Psychological Science*, *16*(2), 145–151. https://doi.org/10.1111/j.0956-7976.2005.00795.x

Theagarajan, R., & Bhanu, B. (2020). Defending black box facial recognition classifiers against adversarial attacks. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, *2020-June*, 3537–3547. https://doi.org/10.1109/CVPRW50498.2020.00414

Toole, A. J. O., & Roark, D. (2006). Predicting Human Performance for Face Recognition. *Face Processing*, *2006*, 293–319. https://doi.org/10.1016/b978-012088452-0/50010-8

Towler, A. (2016). *Match me if you can : Evaluating professional training for facial image comparison* (Issue January). University of New South Wales.

Towler, A., Kemp, R. I., Mike Burton, A., Dunn, J. D., Wayne, T., Moreton, R., & White, D. (2019). Do professional facial image comparison training courses work? *PLoS ONE*, *14*(2), 1–17. https://doi.org/10.1371/journal.pone.0211037

Towler, A., Kemp, R. I., & White, D. (2017). Unfamiliar Face Matching Systems in Applied Settings. In M. Bindemann & A. M. Megreya (Eds.), *Face Processing: Systems, Disorders and Cultural Difficulties* (pp. 21–40). Nova Science Publishers.

Towler, A., White, D., Ballantyne, K., Searston, R. A., Martire, K. A., & Kemp, R. I. (2018). Are Forensic Scientists Experts? *Journal of Applied Research in Memory and Cognition*, *7*(2), 199–208. https://doi.org/10.1016/j.jarmac.2018.03.010

Towler, A., White, D., & Kemp, R. (2021). Can face identification ability be trained? Evidence for two routes to expertise. In M. Bindemann (Ed.), *Forensic Face Matching*. Oxford University Press.

Towler, A., White, D., & Kemp, R. I. (2014). Evaluating training methods for facial image comparison: The face shape strategy does not work. *Perception*, *43*(2–3), 214–218.

https://doi.org/10.1068/p7676

Towler, A., White, D., & Kemp, R. I. (2017). Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied*, *23*(1), 47–58. https://doi.org/10.1037/xap0000108

Tsifouti, A. (2016). *Image usefulness of compressed surveillance footage with different scene contents*. *May*. https://ezp.lib.unimelb.edu.au/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edsble&AN=edsble.690535&site=eds-live&scope=site

Tsushima, Y., & Watanabe, T. (2009). Roles of attention in perceptual learning from perspectives of psychophysics and animal learning. *Learning and Behavior*, *37*(2), 126–132. https://doi.org/10.3758/LB.37.2.126

Tummon, H. M., Allen, J., & Bindemann, M. (2019). Facial Identification at a Virtual Reality Airport. *I-Perception*, *10*(4). https://doi.org/10.1177/2041669519863077

Turk, M. A., & Pentland, A. P. (1991). *Face recognition using eigenfaces* (pp. 586–591). https://doi.org/10.5120/20740-3119

Ulery, B. T., Hicklin, R. A., Buscaglia, J. A., & Roberts, M. A. (2012). Repeatability and reproducibility of decisions by latent fingerprint examiners. *PLoS ONE*, *7*(3). https://doi.org/10.1371/journal.pone.0032800

Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2011). Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(19), 7733–7738. https://doi.org/10.1073/pnas.1018707108

Ulery, B. T., Hicklin, R. A., Kiebuzinski, G. I., Roberts, M. A., & Buscaglia, J. A. (2013). Understanding the sufficiency of information for latent fingerprint value determinations. *Forensic Science International*, *230*(1–3), 99–106. https://doi.org/10.1016/j.forsciint.2013.01.012

Ulery, B. T., Hicklin, R. A., Roberts, M. A., & Buscaglia, J. A. (2014). Measuring what latent fingerprint examiners consider sufficient information for individualization determinations. *PLoS ONE*, *9*(11). https://doi.org/10.1371/journal.pone.0110179

Ulery, B. T., Hicklin, R. A., Roberts, M. A., & Buscaglia, J. A. (2015). Changes in latent fingerprint examiners' markup between analysis and comparison. *Forensic Science International*, *247*(1), 54–61. https://doi.org/10.1016/j.forsciint.2014.11.021

Ulery, B. T., Hicklin, R. A., Roberts, M. A., & Buscaglia, J. A. (2016). Interexaminer variation of minutia markup on latent fingerprints. *Forensic Science International*, *264*, 89–99. https://doi.org/10.1016/j.forsciint.2016.03.014

Ulery, B. T., Hicklin, R. A., Roberts, M. A., & Buscaglia, J. A. (2017). Factors associated with latent fingerprint exclusion determinations. *Forensic Science International*, *275*, 65–75. https://doi.org/10.1016/j.forsciint.2017.02.011

Valentine, T. (1991). A Unified Account of the Effects of Distinctiveness, Inversion, and

Race in Face Recognition. *The Quarterly Journal of Experimental Psychology Section A*, *43*(2), 161–204. https://doi.org/10.1080/14640749108400966

White, D., Burton, A. M., Jenkins, R., & Kemp, R. I. (2014). Redesigning photo-ID to improve unfamiliar face matching performance. *Journal of Experimental Psychology. Applied*, *20*(2), 166–173. https://doi.org/10.1037/xap0000009

White, D., Burton, A. M., Kemp, R. I., & Jenkins, R. (2013). Crowd effects in unfamiliar face matching. *Applied Cognitive Psychology*, *27*(6), 769–777. https://doi.org/10.1002/acp.2971

White, D., Dunn, J. D., Schmid, A. C., & Kemp, R. I. (2015). Error rates in users of automatic face recognition software. *PLoS ONE*, *10*(10). https://doi.org/10.1371/journal.pone.0139827

White, D., Kemp, R. I., Jenkins, R., & Burton, A. M. (2014). Feedback training for facial image comparison. *Psychonomic Bulletin {&} Review*, *21*(1), 100–106. https://doi.org/10.3758/s13423-013-0475-3

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PLoS ONE*, *9*(8). https://doi.org/10.1371/journal.pone.0103510

White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1814), 20151292. https://doi.org/10.1098/rspb.2015.1292

White, D., Towler, A., & Kemp, R. I. (2021). Understanding professional expertise in unfamiliar face matching. In M. Bindemann (Ed.), *Forensic Face Matching: Research and practice*. Oxford University Press.

Wilkinson, C., & Evans, R. (2009). Are facial image analysis experts any better than the general public at identifying individuals from CCTV images? *Science and Justice*, *49*(3), 191–196. https://doi.org/10.1016/j.scijus.2008.10.011

Wilmer, J. B. (2017). Individual Differences in Face Recognition: A Decade of Discovery. *Current Directions in Psychological Science*, *26*(3), 225–230. https://doi.org/10.1177/0963721417710693

Wirth, B. E., & Carbon, C.-C. (2017). An easy game for frauds? Effects of professional experience and time pressure on passport-matching performance. *Journal of Experimental Psychology: Applied*, *23*(2), 138–157. https://doi.org/10.1037/xap0000114

Wixted, J. T., Don Read, J., & Stephen Lindsay, D. (2016). The Effect of Retention Interval on the Eyewitness Identification Confidence–Accuracy Relationship. *Journal of Applied Research in Memory and Cognition*, *5*(2), 192–203. https://doi.org/10.1016/j.jarmac.2016.04.006

Wixted, J. T., & Wells, G. L. (2017). The Relationship Between Eyewitness Confidence and Identification Accuracy: A New Synthesis. *Psychological Science in the Public Interest*, *18*(1), 10–65. https://doi.org/10.1177/1529100616686966

Woodhead, M. M., Baddeley, a. D., & Simmonds, D. C. V. (1979). On Training People to Recognize Faces. *Ergonomics*, *22*(3), 333–343. https://doi.org/10.1080/00140137908924617

Yan, X., Young, A. W., & Andrews, T. J. (2017). The automaticity of face perception is influenced by familiarity. *Attention, Perception, and Psychophysics*, *79*(7), 2202–2211. https://doi.org/10.3758/s13414-017-1362-1

Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. *Conference on Human Factors in Computing Systems - Proceedings*, *January*. https://doi.org/10.1145/3313831.3376301

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141–145.

Young, A. W., & Burton, A. M. (2017). Recognizing Faces. *Current Directions in Psychological Science*, *26*(3), 212–217. https://doi.org/10.1177/0963721416688114

Young, A. W., & Burton, A. M. (2018). Are We Face Experts? *Trends in Cognitive Sciences*, *22*(2), 100–110. https://doi.org/10.1016/j.tics.2017.11.007

Young, A. W., & Noyes, E. (2019). We need to talk about super-recognizers Invited commentary on: Ramon, M., Bobak, A. K., & White, D. Super-recognizers: From the lab to the world and back again. British Journal of Psychology. *British Journal of Psychology*, *110*(3), 492–494. https://doi.org/10.1111/bjop.12395

Zhang, J., & Mueller, S. T. (2005). A note on ROC analysis and non-parametric estimate of sensitivity. *Psychometrika*, *70*(1), 203–212. https://doi.org/10.1007/s11336-003-1119-8

Zhou, X., & Jenkins, R. (2020). Dunning–Kruger effects in face perception. *Cognition*, *203*(May). https://doi.org/10.1016/j.cognition.2020.104345

# *Appendix A – Training survey*

**Facial Comparison Training Survey**

**Introduction** Thank you for considering to take part in this survey. Below is some information about the aims of the survey that will help you decide whether you wish to complete it. Even if you start the survey, you can decide to stop at any time by closing your browser and we will not use any of the information that you give us.

This survey is part of research being carried out by the Centre for Policing Research and Learning at The Open University and the Forensic Organisation of Scientific Area Committees (OSAC).

Your individual responses will NOT be given to police forces/organisations, to any other part of the criminal justice system or shared with anyone outside the research team. The information will be stored completely anonymously (we will not ask for your name for example) and only summaries of the information will ever be presented (for example the 'average' or most common response).

The aim of the survey is to find out more about current practices in facial comparison training in applied settings (e.g. policing, forensics, security and border control). The survey will ask questions on the type of training provided by your agency, what specific topics are covered by the training and how the training is delivered. The results of this survey will be used to direct future research into effective training strategies in facial identification and the production of good practice guidance and standards in the field.

If you would like to take part in the survey, then please click on the arrow button below. This will indicate your consent that you wish to take part and start the survey. You can take as much time as you like to complete the survey, but it will probably take between 10 and 15 minutes.

Q30 In which country is your agency/organisation based?

_____

1 Please enter your agency/organisation name. If you cannot, or would prefer not to provide this information please leave blank.

_____

Q34 Please enter your department name. If you cannot, or would prefer not to provide this information please leave blank.

3 What types of facial comparison does your agency undertake/provide training in?

**Facial Examination** (The task of facial examination includes, but is not limited to, a rigorous morphological analysis, comparison, and evaluation of controlled and uncontrolled images for the purpose of effecting a conclusion. Examiners in this situation have to draw on a larger foundation of knowledge, skill, and ability to accurately reach their conclusions. Examiners should have an in-depth knowledge of the application of available tools and be able to articulate the scientific and legal basis for the expression of conclusions. Facial examination requires an advanced level of training to include an expanded set of knowledge, skills, and abilities above facial assessment and review.) (1)

**Facial Review** (The task of facial review includes, but is not limited to, the use of a facial recognition system to review one-to-many galleries. This task may also include applications involving high volume throughput or escalations from facial assessment. Reviewers require a basic level of training to acquire general knowledge and comprehension of the technology and major elements of the facial comparison discipline and training in the use of available tools. For example, an intelligence analyst will conduct a one-to-many search of a controlled image against a database of controlled images.) (4)

4 What types of material do you receive for comparison?

Still images (e.g. photographs or images taken from video) (1)

Video (2)

Candidate lists from an automated recognition system (3)

Live subjects (4)

Other (5)

5 Who supplies your agency's facial comparison training?

Internally within your agency (1)

Externally from another agency (2)

Externally from a commercial provider (3)

Other (4)

We supply training to other agencies (6)

Our agency does not provide training (5)

6 Does your agency plan to provide facial comparison training in the future?

Yes (1)

No (2)

Q38 What is the duration of training new practitioners receive in your agency (i.e. before they can undertake independent comparisons)?

| | Less than 1 day (1) | 1 day (2) | 2 to 5 days (3) | 2 to 4 weeks (4) | 1 to 6 months (5) | 6 to 12 months (6) | 1 to 5 years (7) | 5 + years (8) |
|---|---|---|---|---|---|---|---|---|

9 What type of training do your reviewers receive?

- Online training (1)
- Independent learning (2)
- Instructor driven seminars (3)
- One-to-one mentoring (4)
- Other (5)

8 What type of training do your examiners receive?

- Online training (1)
- Independent learning (2)
- Instructor driven seminars (3)
- One-to-one mentoring (4)
- Other (5)

11 What topics are covered by the training?

- Facial anatomy (1)
- Image capture and recording (2)
- Image processing (3)
- Methods of comparison (4)
- Other (5)

12 What facial anatomy topics are covered by the training?

| |
|---|
| Face shape (1) |
| Eyes (2) |
| Ears (3) |
| Nose (4) |
| Mouth (5) |
| Chin and jaw (6) |
| Features of the skin (e.g. scars and marks) (7) |
| Bones of the skull (8) |
| Muscles of the face (9) |
| Creases and lines (10) |
| Facial Expression (11) |
| Effects of ageing (12) |
| Juvenile development (13) |
| Permanence of features (14) |
| Alterations to the face (e.g. piercings, tattoos, cosmetic surgery) (15) |
| Other (16) |

13 Which aspects of image capture and recording are covered by the training?

| |
|---|
| Properties of visible light (1) |
| Properties of non-visible wavelengths (e.g. near infrared, ultraviolet) (2) |
| Digital image capture and camera sensors (3) |
| Impact of lighting and camera exposure (4) |
| Geometric distortions (e.g. perspective, lens distortions) (5) |
| Aspect ratio distortion (6) |
| Pixel resolution (7) |
| Image compression (e.g. JPEG) (8) |
| Video compression (e.g. H.264, MPEG-4) (9) |
| Other (10) |

14 What image processing topics are covered by the training?

| |
|---|
| Brightness and contrast adjustments (1) |
| Rotations and cropping (2) |
| Sharpening and blurring (3) |
| Scaling (4) |
| Colour channel seperation (5) |
| Effects of image processing on facial appearance (6) |
| Other (7) |

15 What methods of comparison are covered by the training?

| |
|---|
| Instruction in the ACE-V framework (analyse, compare, evaluate - verify) (1) |
| Instruction in holistic comparison (i.e. comparing the face as a whole) (2) |
| Limitations of holistic comparison (3) |
| Instruction in morphological feature comparison (i.e. comparing the face feature-by-feature) (4) |
| Limitations of morphological feature comparison (5) |
| Instruction in facial feature classification (i.e. assigning features to categories based on appearance e.g. face shape) (6) |
| Limitations of facial feature classification (7) |
| Instruction in photo anthropometry (i.e measuring/comparing the proportions of the face) (8) |
| Limitations of photo anthropometry (9) |
| Instruction in superimposition (i.e overlaying or blending two facial images) (10) |
| Limitations of superimposition (11) |
| Use of automated facial recognition algorithms (12) |
| Human facial recognition (i.e. recognising a known person) (13) |
| Cognitive bias (14) |
| Own-race effects (18) |
| Evaluating comparison findings (15) |
| Peer-review and independent verification (16) |
| Other (17) |

16 Does the training include any of the following exercises?

| |
|---|
| One-to-one facial comparisons (1) |
| One-to-many facial comparisons (3) |
| Feedback on comparison responses (14) |
| Comparison of specific facial features (5) |
| Working on comparisons in pairs or small groups (7) |
| Facial comparisons using a facial feature list (9) |
| Enhancement of images for comparison (10) |
| Testing comparison ability prior to training (11) |
| Testing comparison ability after training (12) |
| Other (13) |

Q28 Any other comments - e.g. clarification on any of your responses or additional information not covered by the survey

*Appendix B – Face matching training course overview*

# Facial Comparison Awareness Course

**Course Content and Learning Outcomes**

| **Module 1 - Introduction to Facial Identification** | |
|---|---|
| **Content** | ▪ Course introduction<br>▪ Types of facial identification<br>▪ Comparison vs. recognition<br>▪ Facial comparison as evidence or intelligence |
| **Exercises** | ▪ Group recognition task<br>▪ Group comparison task |
| **Learning Outcomes** | ▪ Overview of structure and content of course<br>▪ Awareness of the different types of facial identification<br>▪ Understand the difference between comparison and recognition<br>▪ Understand the use of facial identification as an intelligence tool<br>▪ Awareness of the legal requirements for evidential facial identification |

| **Module 2 - Morphological Comparison** | |
|---|---|
| **Content** | ▪ Overview of facial feature based morphological comparison<br>▪ Muscles of the face<br>▪ Facial feature anatomy<br>o Eyes<br>o Ears<br>o Nose<br>o Mouth<br>▪ Scars, marks and blemishes<br>▪ Creases and wrinkles<br>▪ Hair growth patterns |
| **Exercises** | ▪ One-to-one comparisons<br>▪ Individual feature comparisons |
| **Learning Outcomes** | ▪ Learn how to compare faces on a feature-by-feature basis<br>▪ Awareness of the strengths and limitations of morphological comparison<br>▪ Understand the concepts of class, sub-class and fine feature detail<br>▪ Understand the significance of different types of features for comparison |

| Module 3 - Imaging and Environmental Factors | |
|---|---|
| **Content** | ▪ Basics of digital imaging<br>▪ Image quality issues<br>▪ Lighting conditions<br>▪ Pose and camera angle |
| **Exercises** | ▪ One-to-one comparisons using low quality images<br>▪ Comparing images captured at different camera angles and poses |
| **Learning Outcomes** | ▪ Awareness of how digital images are created and processed<br>▪ Understand the impact of image quality on facial appearance<br>▪ Understand the impact of lighting conditions on facial appearance<br>▪ Understand how features vary due to pose and camera angle |

| Module 4 - Automated Facial Recognition (AFR) | |
|---|---|
| **Content** | ▪ Overview of automated systems<br>▪ Image quality for machines<br>▪ Reviewing results from algorithms |
| **Exercises** | ▪ One-to-many comparisons |
| **Learning Outcomes** | ▪ Awareness of how AFR systems work<br>▪ Understand how AFR systems are used operationally<br>▪ Awareness of the advantages and weaknesses of AFR systems<br>▪ Understand concept of probe images, enrolment, galleries, match scores. thresholding and candidate lists<br>▪ Understand how to review candidate lists from AFR searches |

| Module 5 - Human Factors | |
|---|---|
| **Content** | ▪ Overview of cognitive factors affecting comparison<br>▪ False positives and false negatives<br>▪ Mitigating bias and reducing errors<br>▪ Confidence vs. accuracy |
| **Exercises** | ▪ One-to-one comparisons using multiple images |
| **Learning Outcomes** | ▪ Awareness of cognitive factors in comparison decisions<br>▪ Understanding different sources of bias and bias mitigation<br>▪ Understanding and managing different sources of error |

| Module 6 - Comparison Processes and Procedures | |
|---|---|
| **Content** | ▪ Applying morphological comparison operationally<br>▪ Decision making and reporting of results<br>▪ Quality assurance processes |
| **Exercises** | ▪ One-to-one comparisons with quality assurance |
| **Learning Outcomes** | ▪ Understand how to apply a morphological approach operationally<br>▪ Understand how to report decisions<br>▪ Understanding of quality assurance processes |

| Module 7 - Facial Growth and Development | |
|---|---|
| **Content** | ▪ Overview of facial growth and development<br>▪ Individual variations in growth and development<br>▪ AFR performance using juvenile images<br>▪ Challenges of juvenile facial comparison |
| **Exercises** | ▪ One-to-many comparisons of juveniles |
| **Learning Outcomes** | ▪ Awareness of processes of facial growth and development<br>▪ Understanding how features vary with age<br>▪ Understanding human and AFR performance for juvenile identification |

# *Appendix C - Face matching trial online consent form*

**Face Comparison Exercises**

The following test consists of a series of facial image pairs. You must compare the facial images and decide whether the images are of the same person (match) or show different people (non-match). You must also provide a level of confidence in your decision from 1 (not at all confident) to 4 (extremely confident). You may go back during the test to change your answers. It is estimated the test should take no longer than an hour. It is recommended that you view the images in full screen mode (press F11 or change the view settings in your browser).

In order to participate in this study please read and accept the consent form below. Your data will only be used in the study if you have accepted the consent form below.

Thank you.

CONSENT FORM - Facial Comparison Study

I understand that my participation in this study will involve two facial comparison tests on different days.

I understand that participation in this study is entirely voluntary.

I understand that I am free to ask any questions at any time. I am free to withdraw without providing a reason, or having to discuss my concerns with the experimenter.

I understand that I may withdraw from the study up to the point that the data is anonymised on 27th February 2019.

I understand that after 27th February 2019 my data will be held anonymously so that it is impossible to trace this information back to me individually. In accordance with the Data Protection Act this information may be retained indefinitely.

I understand that at the end of the study I will be provided with additional information about the purpose of the study.

I have had an opportunity to discuss with the experimenter any questions or concerns I have about the study.

I consent to participate in this study conducted by Reuben Moreton in the Faculty of Social Science, The Open University.

# *Appendix D – Face matching interface*